SAPIENZA
UNIVERSITÀ DI ROMA

Università di Roma La Sapienza
Scuola di dottorato in Scienze Statistiche
Curriculum: Scienze Attuariali

Un approccio Quantile Regression per la tariffazione danni, basato su un modello a due parti.

Tutor
Prof. Massimiliano Menzietti

Candidato
Davide Biancalana
XXIX Ciclo

Febbraio 2017
Anno Accademico 2016/2017

Sommario

1. INTRODUZIONE.

La tesi trae ispirazione dalla Direttiva Solvency II, che amplia e modifica il metodo di calcolo del requisito di capitale di solvibilità per le imprese di assicurazione operanti all'interno dell'Unione Europea. Tale Direttiva introduce l'obbligo per ogni impresa di assicurazione, dopo aver considerato l'insieme dei rischi a cui è esposta e tenendo conto dei rischi specifici sottoscritti, di determinare un requisito di capitale specifico per il rischio di tariffazione (c.d. "Premium Risk") derivante dai contratti da sottoscrivere nell'anno successivo e dai rischi ancora in vigore sui contratti esistenti.

La Direttiva individua come principio generale di calcolo del requisito di capitale il Value at Risk (**di seguito $\boldsymbol{VaR}$**) a livello 99,5%, nell'orizzonte temporale di un anno, della distribuzione della fonte di rischio oggetto di analisi, che nel caso in esame è il danno aggregato; per cui i fondi di cui deve disporre l'impresa per essere solvibile, in modo coerente con la nuova Direttiva, constano della somma di una componente equa o attesa e di una componente di caricamento necessaria a raggiungere il percentile della distribuzione del danno aggregato al livello di probabilità suddetto.

Il fatto che la distribuzione di probabilità di riferimento sia quella del danno aggregato è centrale all'interno del lavoro; quest'ultimo, infatti, mira proprio alla definizione di un criterio coerente di distribuzione del requisito di capitale sulle singole teste assicurate, tenendo conto delle caratteristiche del singolo assicurato, al fine di definire un caricamento individuale coerente con una specifica misura di rischio. Infatti una volta definito un requisito di capitale sul danno aggregato, è naturale pensare che ciascun assicurato contribuirà in modo diverso a tale requisito, poiché diverse sono le sue caratteristiche rispetto ai fattori rappresentativi del rischio.

La definizione di un requisito di capitale, funzione di una misura di rischio del danno aggregato, stabilisce un'analogia evidente con i principi di calcolo del premio noti nella letteratura attuariale (Daboni 1989). La teoria dell'utilità considera il fatto che l'assicurato, soggetto economico avverso al rischio, accetti la non equità del contratto assicurativo, nonostante il caricamento imposto dall'assicuratore e purchè esso non sia troppo elevato (tale caricamento può anche essere giustificato con il criterio della probabilità di rovina).

Un'impresa di assicurazione, come noto, al fine di stimare gli esborsi a cui andrà incontro nel periodo di validità della tariffa, considera il valore atteso dell'esborso complessivo più una componente di caricamento la cui determinazione è riconducibile, in linea generale, a un percentile della distribuzione del danno aggregato (maggiore del valore atteso) a un certo livello di probabilità, come nel caso Solvency II.

Considerando il problema da un punto di vista della tariffazione, dunque, si vuole definire nel presente elaborato, un principio metodologico che consenta di identificare il premio individuale (e quindi il relativo caricamento) che dovrà essere richiesto a ciascun assicurato, anche in base della sua specifica rischiosità, e che permetta di coprire l'esborso complessivo (comprensivo di caricamento) stimato dall'impresa (c.d. fabbisogno puro).

Nell'ambito della tecnica attuariale la determinazione di un premio individuale funzione di caratteristiche di rischio osservabili a priori sul singolo assicurato (dette variabili tariffarie) è definita personalizzazione a

priori. La tecnica maggiormente utilizzata, ad oggi, è quella dei Generalized Linear Model (di seguito **GLM**) (Nelder, Wedderburn 1972); tali modelli regressivi, a partire da un'ipotesi distributiva della variabile risposta (l'unico vincolo è che tale distribuzione appartenga alla famiglia esponenziale), nell'uso comune, mirano a definire una stima del valore atteso condizionato del danno, quindi una stima del premio equo individuale.

Utilizzando i GLM per definire una stima del valore atteso condizionato, si ottiene in output una sequenza di premi equi personalizzati, la cui somma (per la linearità della media) permette la copertura del fabbisogno al netto del caricamento, ovvero del valore atteso del danno aggregato. La personalizzazione a priori effettuata attraverso il GLM, crea una differenza tra il fabbisogno puro (obiettivo dell'impresa) e i premi effettivamente incassati, in quanto la componente di caricamento del danno aggregato non sarebbe coperta. A oggi, nella pratica, per ottenere l'equilibrio tra le entrate e le uscite dell'impresa, la differenza tra il fabbisogno puro e la somma dei premi equi individuali è sanata distribuendo la differenza stessa, in modo proporzionale su tutti gli assicurati indipendentemente dalle loro caratteristiche, ovvero moltiplicando ogni premio equo individuale per la stessa costante di proporzionalità maggiore di uno.

Risulta dunque evidente che i GLM non definiscono un criterio coerente di capital allocation del fabbisogno puro sulle teste assicurate in funzione delle loro caratteristiche, ma definisce "soltanto" un'allocazione della componente equa (attesa) del fabbisogno stesso.

Si fa notare che i GLM, data l'ipotesi di distribuzione della variabile risposta, possono essere utilizzati anche per una stima della distribuzione di probabilità individuale e quindi per la definizione del caricamento individuale. Tuttavia, come sarà mostrato nell'elaborato, tale approccio è nella maggior parte dei casi inefficace a causa delle ipotesi sottostanti le stime dei parametri, e pertanto si è pensato di adottare una tecnica alternativa originale basata sulla Quantile Regression.

Quest'ultima è una tecnica di regressione (Koenker, Basset 1978) che permette la stima del quantile condizionato della variabile risposta in funzione di un certo numero di regressori; la letteratura e gli ambiti di applicazione sono molto vaste in ambito statistico, mentre in ambito attuariale, e in particolar modo nella tariffazione, l'utilizzo è molto più limitato e riconducibile al solo lavoro di Kudryavtsev (2009). La stima del quantile condizionato suggerisce, quindi, la possibilità di definire, direttamente come output della regressione, il premio puro individuale.

Il passaggio dal GLM alla Quantile Regression permetterebbe un'allocazione, in funzione delle caratteristiche dell'assicurato, non solo del valore atteso del danno aggregato, ma anche della componente di caricamento (ovvero del fabbisogno puro), senza dover ricorrere a una ripartizione proporzionale di quest'ultima.

Nel corso della tesi verranno tuttavia segnalati alcuni limiti dovuti all'introduzione della Quantile Regression in ambito tariffario; passare dalla stima dei valori attesi condizionati (output del GLM), a quella dei quantili condizionati (output della Quantile Regression), comporta l'introduzione di molte proprietà vantaggiose tipiche dei quantili (robustezza, invarianza per trasformazioni monotone), ma anche la perdita delle buone proprietà del valore atteso (linearità, omogeneità, possibilità di fattorizzare la media di variabili casuali indipendenti).

Nel presente elaborato oltre ad introdurre la Quantile Regression nella tariffazione, si è posto l'obiettivo di costruire un impianto tariffario che mirasse ad allocare il percentile del danno aggregato sulle teste assicurate e ad eliminare i problemi dovuti all'assenza di proprietà vantaggiose della media non presenti nei quantili. A tal fine, strumenti fondamentali nella realizzazione del nuovo impianto sono i modelli a due parti (Duan 1984) e la funzione di Laplace Asimmetrica.

Nel capitolo 2 sono introdotte la Direttiva Solvency II e la personalizzazione del premio, che sono concetti fondamentali nella definizione degli obiettivi della tesi; nel capitolo 3 sono definite dal punto di vista teorico alcune tecniche univariate di personalizzazione e i GLM; nel capitolo 4 vi è un introduzione teorica della Quantile Regression; nel capitolo 5 si trova la descrizione del modello attuariale di personalizzazione del premio attraverso i GLM utilizzato nella pratica attuariale e l'introduzione dell'impianto tariffario di tipo Quantile Regression definito in questo lavoro; il tutto si conclude con una parte applicativa nel capitolo 6.

2. IL CONTESTO METODOLOGICO E LA DEFINIZIONE DEGLI OBIETTIVI.

2.1. LA DIRETTIVA SOLVENCY II

La Direttiva Solvency II richiede alle Compagnie operanti nell'Unione Europea di introdurre un requisito di capitale, di seguito "Solvency Capital Requirement" (o "SCR") orientato al rischio, basato su un calcolo prospettico e non inferiore ad un requisito patrimoniale minimo chiamato "Minimum Capital Requirement" (o "MCR").
L'impianto metodologico richiesto da Solvency II necessita della definizione delle seguenti grandezze:

- Un'adeguata misura di rischio;
- Il livello di confidenza a cui riferire la misura di rischio stessa;
- L' orizzonte temporale di valutazione;

Il legislatore richiede alle Compagnie di disporre di un adeguato livello patrimoniale volto a garantire che le imprese siano ancora in grado di onorare i loro obblighi nei confronti dei contraenti e dei beneficiari nei 12 mesi successivi con un livello di probabilità del 99,5%.

La Direttiva Solvency II, inoltre, consente alle Compagnie di calcolare il requisito scegliendo tra cinque possibili metodi che sono proporzionali alla natura, alle dimensioni ed alla complessità dei rischi che le Compagnie stesse dovranno misurare:

- Modello Interno Completo (IRM);
- Formula Standard e Modello Interno Parziale (PIRM);
- Formula Standard con Parametri Specifici dell'Impresa (USP);
- Formula Standard (SF);
- Semplificazioni;

Le Compagnie che decidono di effettuare valutazioni con un modello interno completo o parziale dovranno giustificare il metodo all'Autorità di Vigilanza e dovranno effettuare una scelta su ognuna delle tre grandezze sopra definite per tutti i rischi quantificabili elencati nell'Articolo 101 della Direttiva:

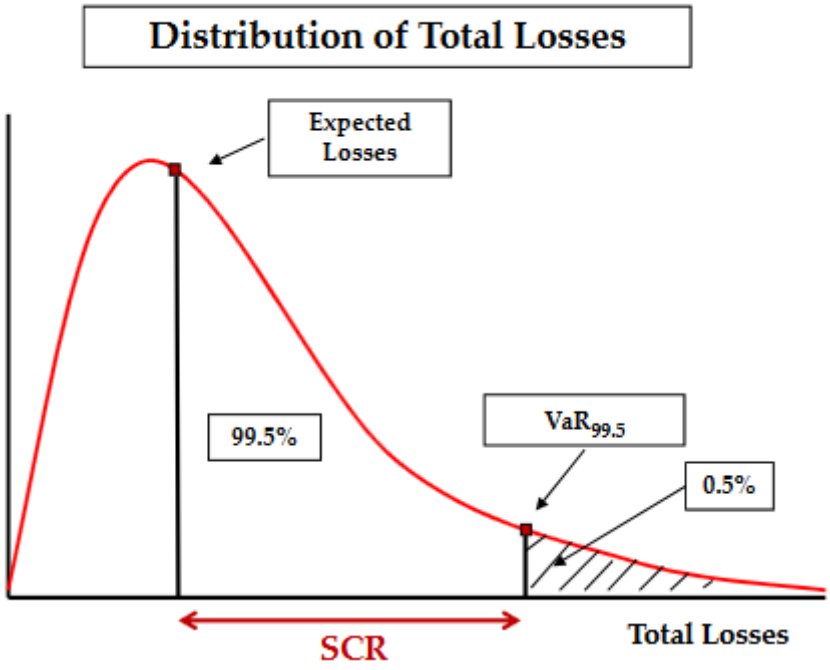
- Il rischio di sottoscrizione per l'assicurazione non vita;
- Il rischio di sottoscrizione per l'assicurazione vita;
- Il rischio di sottoscrizione per l'assicurazione malattia;
- Il rischio di mercato;
- Il rischio di credito
- Il rischio operativo (include i rischi giuridici, ma non quelli derivanti da decisioni strategiche e/o i rischi di reputazione che, invece, sono valutati nel Pilastro II).

2.2. DEFINIZIONE DI SCR E LA FORMULA STANDARD

Analogamente al Margine di Solvibilità richiesto dalla normativa precedente, anche la Direttiva Solvency II prevede una Formula Standard per la determinazione dei requisiti di capitale.

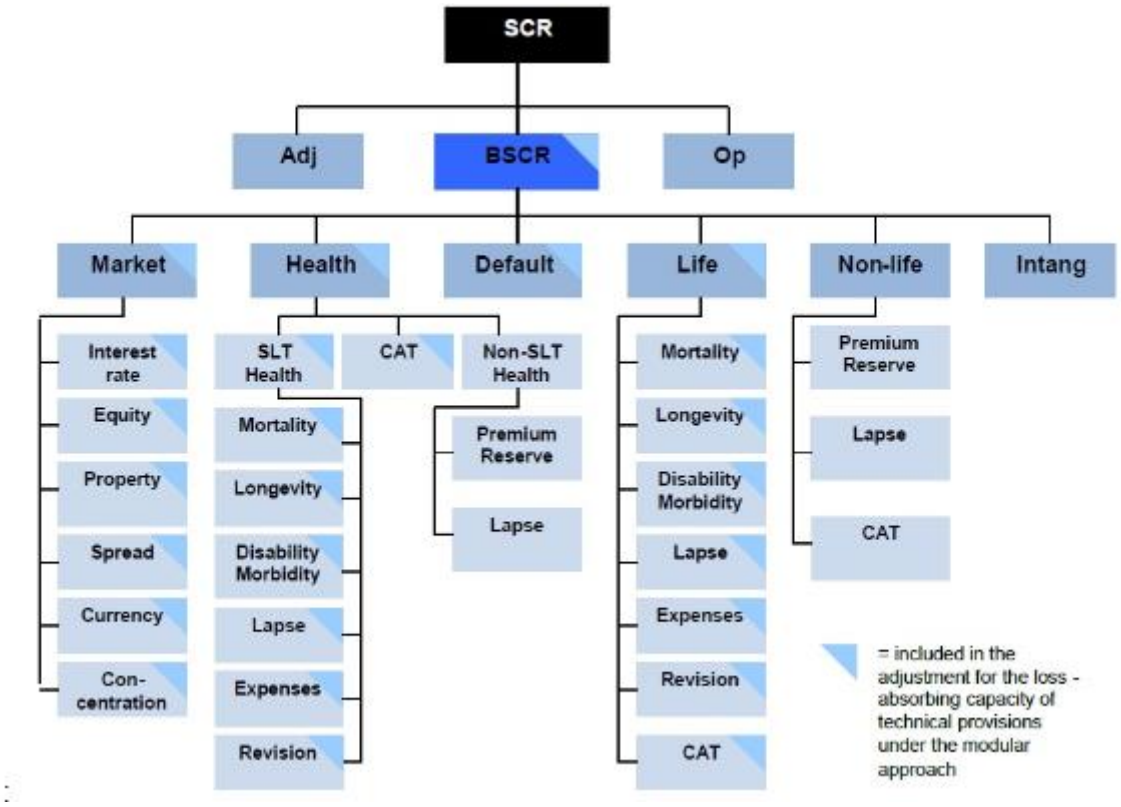
La misura di rischio scelta per la determinazione del SCR secondo la SF è il Value at Risk (di seguito"VaR") calibrato ad un livello di confidenza pari al 99.5% in orizzonte temporale annuale. Tale requisito deve tenere conto di ogni tecnica di mitigazione del rischio (ad es. riassicurazione, securitisation, ecc.) e, a seconda della natura dei rischi già sintetizzati, è determinato secondo un approccio basato su fattori moltiplicativi (c.d. factor based approach) o secondo un approccio per scenario (c.d. scenario testing).

Figura 1: Definizione SCR



In base alla formula standard, l'SCR globale è calcolato aggregando gli SCR derivanti dai sei rischi quantificabili definiti pocanzi in base al seguente schema:

Figura 2: Struttura generale del SCR secondo la SF



Il SCR, pertanto, è pari a

$$\text{SCR} = \text{BSCR} + \text{Adj} + \text{SCR}_{\text{op}}$$

Calcolato il requisito di capitale per i rischi operativi ("Op") e gli aggiustamenti per la capacità di assorbimento delle perdite delle riserve tecniche e delle imposte differite ("Adj"), il Basic Solvency Capital Requirement (di seguito "BSCR") è determinato aggregando i requisiti di capitale di ognuna delle sei categorie di rischio elencate nella Figura 3:

$$BSCR = \sqrt{\sum_{ij} Corr_{ij} \times SCR_i \times SCR_j} + SCR_{intangible}$$

Se da un lato tra gli Adj, il Op ed il BSCR il legislatore ha previsto ipotesi di indipendenza nella SF,dall'altro occorre rilevare come una delle maggiori novità apportate dalla Direttiva è la definizione di una matrice di correlazione lineare ($Corr_{ij}$) per la loro aggregazione, ad eccezione del SCR per i rischi connessi agli attivi immateriali ($SCR_{intangible}$).

Figura 3: Matrice di correlazione del SCR secondo la SF

j i	Market	Default	Life	Health	Non-life
Market	1				
Default	0.25	1			
Life	0.25	0.25	1		
Health	0.25	0.25	0.25	1	
Non-life	0.25	0.5	0	0	1

2.3. LA FORMULA STANDARD PER IL NON LIFE UNDERWIRTING RISK

Il Non – Life Underwriting Risk (di seguito "NLUw"), parte del BSCR definito nel paragrafo precedente, assume particolare rilevanza rispetto all'elaborato poiché è il modulo che contiene anche il Premium Risk.
Nelle specifiche tecniche più recenti ed ai sensi di quanto descritto all'interno del quinto studio di impatto quantitativo (di seguito "QIS5") presentato al mercato europeo nel 2011, il NLUw è il rischio proveniente dalle obbligazioni derivanti dalla sottoscrizione di ogni tipologia di rischio dei rami Danni. Tale modulo, inoltre, comprende i rischi derivanti dai processi in atto per la gestione di tale business e dall'incertezza che nasce dal potenziale diritto di esercizio delle opzioni dei contraenti, quali, ad esempio, il rinnovo o l'abbandono del contratto.
Questo rischio si articola in tre sotto moduli:

- NL_{pr}: Non – Life premium and reserve risk;
- NL_{lapse}: Non – Life lapse risk;
- NL_{CAT}: Non – Life catastrophe risk;

Rimandando al prossimo paragrafo per la valutazione del NL_{pr}, occorre chiarire, che il NL_{lapse} è il requisito di capitale utile a far fronte all'errata valutazione a priori delle opzioni implicite dei contratti, tra le quali si menzionano, a titolo esemplificativo, l'abbandono prima della scadenza del contratto e il rinnovo a condizione analoghe a quelle precedenti.

Il NL_{CAT}, invece, valuta il capitale economico da accantonare per far fronte agli eventi a bassa frequenza e ad alto costo medio, derivanti da eventi naturali, da riassicurazione non proporzionale del property, da eventi indotti dall'uomo, nonché per le altre eventuali garanzie catastrofali descritte in polizza.
Definita la seguente matrice di correlazione (CorrNL).
Il NL_{lapse} e il NL_{CAT} , che vengono calcolati secondo un approccio per scenario, sono esclusi dal perimetro di tale tesi di ricerca.

Figura 4: matrice di correlazione della NLUw secondo la formula standard

$CorrNL$	NL_{pr}	NL_{lapse}	NL_{CAT}
NL_{pr}	1		
NL_{lapse}	0	1	
NL_{CAT}	0.25	0	1

Il requisito di capitale per tale modulo secondo la SF dovrà essere calcolato nel seguente modo:

$$SCR_{NL} = \sqrt{\sum_{rc} CorrNL_{rc} \times SCR_r \times SCR_c}$$

2.3.1. IL REQUISITO DI CAPITALE PER IL PREMIUM E IL RESERVE RISK

Dal quinto studio di impatto quantitativo, all'interno di questo sotto-modulo, l'EIOPA ha espresso la volontà di unire all'interno di un'unica formula il calcolo del requisito di capitale del Premium Risk e del Reserve Risk.

$$NL_{pr} = 3 \cdot \sigma \cdot V$$

Dove:

- σ rappresenta la deviazione standard combinata per il premium e per il Reserve risk
- V, invece, rappresenta la misura di volume combinata del Premium e Reserve Risk, eventualmente corretta per l'effetto di diversificazione (geografico) dei rischi sottoscritti.

Sia la deviazione standard che il volume sono dapprima calcolati all'interno di segmenti/Line of Business ("LoB") tra Premium e Reserve Risk e successivamente aggregati.
La misura di volume è calcolata come indicato nella formula (b)

$$\begin{cases} V = \sum_s V_s \\ V_s = \left(V_{prem,s} + V_{res,s}\right) + 0{,}75 + 0{,}25 + DIV_s \end{cases}$$

Dove:

- DIV_s è l'indice di Herfindal di diversificazione geografica;
- $V_{prem,s}$ è la misura di volume per il Premium Risk nella s-esima LoB;
- $V_{res,s}$ è la misura di volume per il Reserve Risk nella s-esima LoB.

In particolare, per il Reserve Risk, tale misura di volume deve essere pari alla best estimate delle riserve sinistri di ogni segmento e, pertanto, potrà essere inferiore all'ammontare recuperabile dai contratti di riassicurazione e da alcuni veicoli per uso speciale.

Il volume del Premium Risk e, conseguentemente, dei premi di ogni segmento, sarà calcolato come la somma del valore attuale dei premi netti che la Compagnia si attende di incassare per tale LoB oltre l'anno successivo da contratti esistenti ($FP_{(existing,s)}$), dal valore attuale dei premi netti che saranno sottoscritti nei futuri 12 mesi, al netto anche dei premi incassati nei 12 mesi successivi alla data di valutazione ($FP_{(future,s)}$), e dal massimo tra i premi di competenza netti stimati per l'anno successivo all'istante di valutazione (P_s) ed i premi emessi netti dell'anno terminato al momento della valutazione ($P_{(last,s)}$):

$$V_{prem,s} = max\left(P_s; P_{(last,s)}\right) + FP_{(existing,s)} + FP_{(future,s)}$$

Nel caso della deviazione standard, la seguente formula di aggregazione tra i segmenti/LoB tiene conto anche delle correlazioni lineari indicate nella matrice CorrS riportata di seguito (v. Figura 5):

$$\sigma = \sqrt{\sum_{s,t} CorrS_{s,t} \cdot \sigma_s \cdot V_s \cdot \sigma_t \cdot V_t}$$

Figura 5: matrice di correlazione della deviazione standard tra Lob nella SF

CorrS	1	2	3	4	5	6	7	8	9	10	11	12
1: Motor vehicle liability	1											
2: Other motor	0,5	1										
3: MAT	0,5	0,25	1									
4: Fire	0,25	0,25	0,25	1								
5: 3rd party liability	0,5	0,25	0,25	0,25	1							
6: Credit	0,25	0,25	0,25	0,25	0,5	1						
7: Legal exp.	0,5	0,5	0,25	0,25	0,5	0,5	1					
8: Assistance	0,25	0,5	0,5	0,5	0,25	0,25	0,25	1				
9: Miscellaneous.	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	1			
10:Np reins. (casualty)	0,25	0,25	0,25	0,25	0,5	0,5	0,5	0,25	0,25	1		
11:Np reins. (MAT)	0,25	0,25	0,5	0,5	0,25	0,25	0,25	0,25	0,5	0,25	1	
12:Np reins. (property)	0,25	0,25	0,25	0,5	0,25	0,25	0,25	0,5	0,25	0,25	0,25	1

La volatilità calcolata a livello di segmento/LoB, assumendo un coefficiente di correlazione α pari al 50% tra Premium e Reserve Risk, è descritta di seguito:

$$\sigma_s = \frac{\sqrt{\left(\sigma_{(prem,s)} \cdot V_{(prem,s)}\right)^2 + \sigma_{(prem,s)} \cdot V_{(prem,s)} \cdot \sigma_{(res,s)} \cdot V_{(res,s)} + \left(\sigma_{(res,s)} \cdot V_{(res,s)}\right)^2}}{V_{(res,s)} + V_{(prem,s)}}$$

Per poter calcolare il requisito di capitale oggetto di questo paragrafo, resta da definire la deviazione standard per il Premium ed il Reserve Risk, sempre diversificata per LoB.

Nel caso in cui una generica Compagnia avesse deciso di calcolare questo requisito di capitale seguendo esclusivamente la SF, il suddetto parametro di volatilità deve essere selezionato dalla tabella indicate di seguito:

Figura 6:Volatility per premium e reserve risk per LoB

Segment	$\sigma_{(prem,s)}$ *Gross Reins.*	$\sigma_{(res,s)}$ *Net Reins.*
1. MTPL and proportional reinsurance	10%	9%
2. Other Motor insurance and proportional reinsurance	8%	8%
3. MAT and proportional reinsurance	15%	11%
4. Fire insurance and proportional reinsurance	8%	10%
5. TPL and proportional reinsurance	14%	11%
6. Credit insurance and proportional reinsurance	12%	19%
7. Legal expenses ins. and proportional reins.	7%	12%
8. Assistance insurance and proportional reins.	9%	20%
9. Miscellaneous ins. and proportional reins.	13%	20%
10. NP reins (cas)	17%	20%
11. NP reins (MAT)	17%	20%
12. NP reins (prop)	17%	20%

2.4. DA SOLVENCY II AL PREMIO ASSICURATIVO

La tesi, come specificato nell'introduzione, è incentrata sul Premium Risk, ovvero il rischio derivante da scostamenti sfavorevoli, nel momento di accadimento, nella frequenza e nella gravità degli eventi assicurati, rispetto a quanto previsto nel calcolo del premio. La SF definisce le volatilità tipiche delle singole LoBs (Figura 6), e di fatto il calcolo di requisiti di capitale per ciascuna LoB, che poi vengono aggregati attraverso la matrice di correlazione in Figura 5.

Il concetto fondamentale, che rappresenta la principale novità introdotta da Solvency II, è che non è più sufficiente limitarsi alla stima del valore atteso del danno aggregato, ossia la best estimate (di seguito BE), o di alcuni indici di dispersione (varianza, asimmetria, curtosi), ma è necessario disporre delle caratteristiche dell'intera distribuzione di probabilità (percentili o forma dell'intera distribuzione); ad esempio, la Direttiva definisce un radicale cambiamento anche nella logica del calcolo delle riserve tecniche sostituendo la prudenzialità implicita nelle riserve con la prudenzialità espressa dal requisito di capitale a protezione dei rischi. Le riserve tecniche sono quindi calcolate come somma di best estimate (di seguito BE) e risk margin (di seguito RM). La best estimate è la media dei flussi di cassa futuri attualizzati, il risk margin viene calcolato come il costo della costituzione di un ammontare di fondi propri a copertura del requisito di capitale.

Ponendo l'attenzione sul requisito di capitale per il Premium Risk della singola LoB, si sottolinea che esso deve essere stimato (con FS o con una delle metodologie alternative definite al paragrafo 2.1) in funzione del percentile a livello 99,5% della distribuzione di probabilità del danno aggregato come mostrato in Figura 1.

Formalmente, ipotizzando una collettività di r assicurati, sia $\ddot{Y}$ la variabile casuale " danno aggregato", definita nel seguente modo:

$\ddot{Y} = \sum_{k=1}^{r} Y_k$,

dove con Y_k si identifica il danno per la k-esima testa assicurata.

Considerando la Figura 1, si può affermare che un'impresa di assicurazione è solvibile in termini di Premium Risk secondo Solvency II, se dispone di un ammontare complessivo pari a:

$$VaR_{99,5}[\ddot{Y}] = E[\ddot{Y}] + SCR \qquad (2.4.1)$$

L'idea di valutare le poste attuariali come somma di una componente attesa più un margine di sicurezza è un principio basilare utilizzato ad esempio nell'ambito della tariffazione (c.d pricing) e stabilisce un collegamento forte con i principi Solvency II. Il punto fondamentale è che l'equazione (2.4.1) fa riferimento alla distribuzione del danno aggregato della LoB, per cui il passaggio dal concetto di requisito di capitale Solvency II a quello di caricamento del premio assicurativo, richiede a sua volta il passaggio dal calcolo di un percentile sulla distribuzione del danno aggregato $\ddot{Y}$, al calcolo di un percentile sulla distribuzione del danno individuale Y_k.

2.4.1. PRINCIPI DI CALCOLO DEL PREMIO.

Nella tecnica attuariale esistono diverse configurazioni di premio, infatti si parla di premio equo, premio puro e premio di tariffa o commerciale.

Il punto di partenza per la definizione del premio, che verrà pagato dall'assicurato a fronte della copertura assicurativa, è il premio equo. Esso corrisponde al valore atteso del totale dei risarcimenti aleatori a carico dell'impresa di assicurazione per ogni individuo, durante il periodo assicurato ($E[Y_k]$).

La seconda configurazione è quella del premio puro che ricomprende al suo interno anche un caricamento, non presente nel premio equo. Il caricamento del premio ha il ruolo di limitare eventuali perdite, qualora la gestione del portafoglio di contratti sia negativa a causa di errori di stima o di un inaspettato aumento della sinistrosità.
La teoria del rischio, infatti, evidenzia che, considerando il solo premio equo, non è garantita la stabilità economica di una Compagnia che assume dei rischi di terzi.
Il caricamento, pertanto, non svolge solo il ruolo comune a tutti i settori industriali di remunerazione del rischio sopportato dall'imprenditore e da ogni azionista della Compagnia, ma anche quello di far fronte al rischio che il premio non sia in grado di fronteggiare tutti i costi e le spese.
Esistono diversi criteri per il passaggio da premio equo a premio puro, che nella pratica attuariale vengono definiti "principi di calcolo del premio", la maggior parte dei quali lega il caricamento a delle misure di sintesi della distribuzione di probabilità della variabile Y_k:

1. Principio della varianza: $P[Y_k] = E[Y_k] + \alpha VAR[Y_k]$, $\alpha > 0$ reciproco di un importo
2. Principio dello scarto quadratico medio: $P[Y_k] = E[Y_k] + \beta\sigma[Y_k]$, $\beta > 0$ numero puro
3. Criterio del valore atteso: $P[Y_k] = E[Y_k] + \gamma E[Y_k]$, $\gamma > 0$ numero puro

Si segnalano le seguenti osservazioni:

i. Per il criterio del valore atteso e dello scarto quadratico medio vale il principio dell'omogeneità positiva, ossia $P[aY_k] = aP[Y_k]. a > 0$
ii. Per tutti i criteri, tranne quello del valore atteso, vale la proprietà traslativa: $P[Y_k + C] = P[Y_k] + C$, $C > 0$.
iii. Il criterio della varianza e del valore atteso godono della proprietà additiva, considerando il danno aggregato Y, se le Y_k sono stocasticamente indipendenti: $P[Y_1 + \cdots + Y_r] = P[Y_1] + \cdots + P[Y_r]$
iv. Per il criterio dello scarto quadratico medio e del valore atteso, considerando il danno aggregato Y è possibile effettuare la sua decomposizione di Y in r importi non negativi $Y_1, \ldots\ldots, Y_r$, in modo tale che sia $P[\ddot{Y}] = P[Y_1 + \cdots + Y_r] \leq P[Y_1] + \cdots P[Y_r]$.

Per tutti i criteri di calcolo sopra richiamati, dunque, esiste un caricamento di sicurezza $m[Y_k] > 0$ tale che $P[Y_k] > E[Y_k]$, di conseguenza il premio puro è espresso come somma di due componenti:

$P[Y_k] = E[Y_k] + m[Y_k]$.

La terza configurazione di premio è il cosiddetto premio di tariffa, che si differenzia dal premio puro per l'aggiunta di caricamenti a fronte delle spese di acquisizione, incasso premi e gestione. Per quanto riguarda questo elaborato tale terza componente non verrà presa in considerazione e tutte le considerazioni successive verranno effettuate considerando l'ipotesi di assenza delle tre voci di spesa suddette.

2.4.2. LA FUNDAMENTAL INSURANCE EQUATION (FIE).

Considerando il problema per un'impresa di assicurazioni di voler stimare l'esborso complessivo a cui questa andrà incontro nel periodo di validità della tariffa, si può affermare che tale fabbisogno stimato, per quanto detto nei paragrafi precedenti, dovrà anch'esso essere provvisto di un caricamento (c.d. fabbisogno puro).

Tale fabbisogno puro, nell'obiettivo della definizione di un modello tariffario Solvency II compatibile, verrà identificato come Value at Risk a un certo livello di probabilità θ della distribuzione del danno aggregato, ovvero come $VaR_\theta[\ddot{Y}] = Quant_\theta[\ddot{Y}]$.

Una volta stimato l'esborso complessivo a cui l'impresa andrà incontro nel periodo di validità della tariffa, al fine di realizzare l'equilibrio economico, occorre che i premi incassati dall'impresa coprano esattamente tale fabbisogno puro.

In ambito attuariale la relazione di equilibrio tra entrate e uscite è nota come Fundamental Insurance Equation, (di seguito **FIE)**, che nel caso in esame assume la seguente forma:

$Quant_\theta[\ddot{Y}] = \sum_{k=1}^{r} P[Y_k]$

L'idea alla base del lavoro è quella di utilizzare, ai fini della definizione dei premi puri individuali, la medesima misura di rischio e quindi la ricerca di un livello di probabilità θ^*, tale per cui è possibile definire il premio puro individuale sotto forma di quantile della distribuzione Y_k, ovvero: $P[Y_k] = Quant_{\theta^*}[Y_k]$.

Di conseguenza la FIE può essere riscritta come:

$$Quant_\theta[\ddot{Y}] = \sum_{k=1}^{r} P[Y_k] = \sum_{k=1}^{r} Quant_{\theta^*}[Y_k] = \sum_{k=1}^{r} \{E[Y_k] + m[Y_k]\}$$

Rispetto a Solvecny II, che si limita ad indicare il percentile sul danno aggregato, l'obiettivo che ci si pone è definire un modello di allocazione del fabbisogno puro sulle singole teste assicurate, attraverso una stima coerente del caricamento individuale $m[Y_k]$.

Nell'ambito della tecnica attuariale la determinazione di un premio individuale funzione di caratteristiche di rischio sul singolo assicurato è definita personalizzazione.

2.5. PERSONALIZZAZIONE DEL PREMIO

I principi di calcolo precedentemente descritti, consentono di definire le relazioni fondamentali alla base della tariffazione. Tali relazioni possono essere espresse sia in riferimento al singolo rischio assicurato, piuttosto che ad un collettivo di rischi omogenei. In quest'ultimo caso, si potrebbe anche considerare l'intera collettività assicurata, senza distinzione delle specifiche caratteristiche di rischio, individuando così un premio indifferenziato per tutti gli assicurati, introducendo un meccanismo di piena solidarietà tra gli stessi. Formalmente, tale ultima condizione si tradurrebbe in un'ipotesi di identica distribuzione delle variabili aleatorie $F_{Y_k} = F_Y$, per ogni k.

In tale contesto la FIE si scriverebbe nel seguente modo:

$Quant_\theta[\ddot{Y}] = r \cdot P[Y] = r \cdot \{E[Y] + m[Y]\}$.

Il problema del calcolo del caricamento individuale si traduce nella soluzione di un equazione nell'incognita $m[Y]$ per cui:

$m[Y] = \frac{Quant_\theta[\ddot{Y}] - r \cdot E[Y]}{r}$.

Tuttavia, nella maggior parte dei casi, ad esempio nella Responsabilità Civile Autoveicoli (di seguito **R.C.A.**), l'impresa di assicurazione è interessata a definire un premio individuale che sia funzione di alcune caratteristiche degli individui che l'impresa stessa ritiene particolarmente rappresentative del rischio, i cosiddetti "fattori di rischio". Ovvero l'impresa è interessata a suddividere i suoi assicurati in classi di rischio omogenee (profili di rischio), all'interno delle quali gli assicurati stessi presentino fattori di rischio identici e a cui, dunque, possa essere applicato lo stesso premio.

Ipotizzando che sulla collettività assicurata agiscano due fattori di rischio, con rispettivamente n_1 ed n_2 modalità; se si indicano con r_{ij} il numero di assicurati che presentano contemporaneamente la modalità i e la modalità j, allora sarà:

$r = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} r_{ij}$.

Si fa notare che nel caso in questione si avranno $n_1 \cdot n_2$ classi di rischio, e occorre stimare altrettanti premi puri individuali; la classe di rischio generica (i, j) è composta da r_{ij} assicurati e si ipotizza che ciascuna delle Y_k con $k = 1, \dots\dots, r_{ij}$ sia distribuita come un'unica variabile Y_{ij}.

Indicato con $P_{ij} = P[Y_{ij}]$ il premio puro individuale per un assicurato appartenente alla classe di rischio (i, j), i premi complessivamente incassati dovranno essere tali da rispettare la FIE:

$$Quant_\theta[\ddot{Y}] = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} r_{ij} \cdot P_{ij} \qquad (2.5.1)$$

La definizione dei premi puri individuali P_{ij} in modo tale che sia realizzata la FIE, equivale a definire una ripartizione del fabbisogno puro $(Quant_\theta[\ddot{Y}])$, sulle singole classi di rischio, tenendo conto delle specifiche distribuzioni di probabilità; attraverso l'individuazione di un caricamento legato ad una misura di rischio individuale $m[Y_{ij}]$.

2.6. IL MODELLO DI RISCHIO

Una volta chiarito l'obiettivo della tesi, ovvero il calcolo del premio puro come misura di rischio individuale, occorre scegliere un modello di rischio, vale a dire la struttura matematica attraverso la quale si vuole definire il premio. Poiché nella prassi si suole fare ricorso ad un modello moltiplicativo, si è deciso di seguire la stessa via, per cui per ognuno degli $n_1 * n_2$ profili, si punta a definire il premio personalizzato:

$$P_{ij} = P^{(0)} \cdot \beta_{1i} \cdot \beta_{2j}, \; i = 1,2, \dots\dots, n_1 \quad j = 1,2, \dots\dots\dots\dots, n_2 \qquad (2.6.1)$$

Dove $P^{(0)}$ è detto premio di riferimento e può essere interpretato, sotto particolari condizioni, come il premio medio non personalizzato, mentre β_{1i}, β_{2j} rappresentano le cosiddette "relatività", ovvero esprimono quanto i profili di rischio pagheranno in più o in meno rispetto al premio medio globale.

L'obiettivo è quello di stimare l'intero set di parametri.

Sostituendo la (2.6.1) nella FIE, si ottiene:

$$Quant_{\theta}[\ddot{Y}] = P^{(0)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} r_{ij} \cdot \beta_{1i} \cdot \beta_{2j} \qquad (2.6.2)$$

Tale formula rappresenta allo stesso tempo la determinazione del premio e il vincolo di equilibrio da rispettare.

Una distinzione metodologica sottostante la scelta di un modello di rischio riguarda il tipo di analisi per la stima dei coefficienti. Quest'ultima potrà essere:

- Univariata,
- Multivariata.

Il limite dell'analisi univariata del rischio concerne l'impossibilità di riuscire a selezionare le variabili significative, tenendo conto delle dipendenze tra le variabili esplicative. Questa limitazione, tuttavia, è bilanciata dalla semplicità di questo approccio.

Inoltre ricorrere ad un modello univariato presenta un ulteriore limite: alcune combinazioni tra le modalità delle variabili esplicative potrebbero essere vuote, o talmente poco popolate da creare rumore statistico (o alea) in tutto il modello, e/o compromettere la ricerca della componente sistematica del rischio oggetto di analisi.

Per far fronte a tale ultima limitazione, i modelli tecnici multivariati sono scelti dalla gran parte delle Compagnie a livello globale.

Si farà riferimento, nel capitolo successivo, ad alcune tecniche di stima dei parametri $\boldsymbol{\beta}$, sia univariate che multivariate, destinando particolare attenzione sia alla tecnica dei GLM, poiché è la tecnica maggiormente utilizzata nella pratica, sia soprattutto alla Quantile Regression (di seguito **QR**) che, pur poco utilizzata in ambito attuariale, consente di superare alcune limitazioni delle tecniche GLM e riveste un ruolo essenziale nell'impianto teorico tariffario qui proposto.

3. MODELLI UNIVARIATI E MULTIVARIATI PER LA STIMA DEI COEFFICIENTI TARIFFARI.

3.1. PURE PREMIUM

I modelli univariati per il calcolo dei coefficienti sono molteplici, quello più semplice è l'approccio "pure premium", o delle relatività intuitive.

Si definiscano le seguenti grandezze, in relazione al collettivo di rischi utilizzato per definire la FIE:

p_{ij} è la quota danni osservata per un individuo di profilo i, j

$$p = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} p_{ij}\cdot r_{ij}}{r} \tag{3.1.1}$$

$$r_{i\cdot} = \sum_{j=1}^{n_2} r_{ij} \tag{3.1.2}$$

$$r_{\cdot j} = \sum_{i=1}^{n_1} r_{ij} \tag{3.1.3}$$

$$p_{i\cdot} = \frac{\sum_{j=1}^{n_2} p_{ij}\cdot r_{ij}}{r_{i\cdot}} \tag{3.1.4}$$

$$p_{\cdot j} = \frac{\sum_{i=1}^{n_1} p_{ij}\cdot r_{ij}}{r_{\cdot j}} \tag{3.1.5}$$

Dalle (3.1.1), (3.1.4), (3.1.5) si ottiene la seguente relazione

$$p \cdot r = \sum_{i=1}^{n_1} p_{i\cdot} \cdot r_{i\cdot} = \sum_{j=1}^{n_2} p_{\cdot j} \cdot r_{\cdot j} \tag{3.1.6}$$

p è una stima di $E[Y]$ (ovvero del premio di riferimento), ed è la quota danni osservata sull'intero portafoglio.

Nello stesso caso, $p_{i\cdot}$ è la quota danni osservata per i rischi aventi determinazione i per la prima variabile tariffaria. L'interpretazione per $p_{\cdot j}$ è del tutto analoga.

Indicando la variabile casuale premio individuale con:

$$E(Y_{ij}) = p \cdot \beta_{1i} \cdot \beta_{2j}$$

indichiamo con $\hat{p}_{ij}$ il valore stimato per $E(Y_{ij})$, così definito:

$$\hat{p}_{ij} = \hat{p} \cdot \hat{\beta}_{1i} \cdot \hat{\beta}_{2j}$$

Consideriamo le seguenti stime:

$$\hat{p} = p$$

$$\hat{\beta}_{1i} = \frac{p_{i\cdot}}{p} \tag{3.1.7}$$

$$\hat{\beta}_{2j} = \frac{p_{\cdot j}}{p} \tag{3.1.8}$$

Si ottiene allora

$$\hat{p}_{ij} = \frac{p_{i\cdot}\cdot p_{\cdot j}}{p}$$

Questo metodo di stima è di immediata interpretazione, ma presenta l'inconveniente di essere influenzato dalla distribuzione dei rischi nelle classi. Prendendo ad esempio l'assicurazione R.C.A., se si considerano le due variabili tariffarie "età del conducente" e "provincia", potrebbe accadere che la stima delle relatività per una certa zona di circolazione sia influenzata dalla distribuzione per età di conducente in quella zona. Infatti,

la quota danni $p_{.j}$ osservata nella zona j-esima, e quindi la stima $\hat{\beta}_{2j}$, potrebbe indicare una zona ad elevata rischiosità, mentre ciò può essere dovuto non alle caratteristiche della zona in sé, ma al semplice fatto che vi circolano molti "cattivi" conducenti. Se la distribuzione degli assicurati varia allora sensibilmente nelle diverse zone, le relatività intuitive determinano delle distorsioni nella relazione (ordinamento) di sinistrosità tra le zone stesse. In particolare per un assicurato appartenente ad una classe di rischio elevata e ad una zona in cui circolano molti "cattivi" assicurati, questo fattore di rischio viene di fatto contato due volte. Più precisamente, come vedremo a breve, in tale caso il premio medio relativo alla zona non coincide con l'esborso medio osservato; situazione questa indicata dagli attuari come "assenza di bilanciamento".

Dopo aver realizzato una tariffa, a partire da un insieme di osservazioni, è interessante verificare che essa soddisfi alcune proprietà. Innanzitutto è bene accertare che gli scostamenti tra i valori osservati p_{ij} e quelli stimati $\hat{p}_{ij}$ non siano troppo sensibili; una seconda esigenza, nota come bilanciamento, richiede che per gruppi "numerosi" di assicurati gli esborsi osservati siano uguali agli introiti che si sarebbero ottenuti applicando la tariffa appena stimata. Siano $\hat{p}_{ij}$ le stime ottenute per $E(Y_{ij})$, con $i = 1,..,n_1$ e $j = 1,..,n_2$; si dirà che il metodo di stima adottato verifica il bilanciamento sulle righe se:

$\sum_{j=1}^{n_2} \hat{p}_{ij} \cdot r_{ij} = \sum_{j=1}^{n_2} p_{ij} \cdot r_{ij}$ per ogni $i = 1,..,n_1$. (3.1.9)

Se definiamo $\hat{p}_{i.}$ tale che

$\sum_{j=1}^{n_2} \hat{p}_{ij} \cdot r_{ij} = \hat{p}_{i.} \cdot r_{i.}$

$\hat{p}_{i.}$ ha il significato di premio medio da applicare agli assicurati con modalità i per la prima variabile tariffaria. Sfruttando la (3.1.4), si ottiene la seguente espressione per la condizione di bilanciamento sulle righe: $\hat{p}_{i.} r_{i.} = p_{i.} r_{i.}$ per ogni $i = 1,..,n_1$;

ovvero:

$\hat{p}_{i.} = p_{i.}$ per ogni $i = 1,..,n_1$.

Quindi la condizione di bilanciamento sulle righe richiede che i premi medi uguaglino gli esborsi medi osservati per gli assicurati aventi la stessa determinazione per la prima variabile tariffaria. Analogamente, diremo che il metodo tariffario verifica il bilanciamento sulle colonne se:

$\sum_{i=1}^{n_1} \hat{p}_{ij} \cdot r_{ij} = \sum_{i=1}^{n_1} p_{ij} \cdot r_{ij}$, per ogni $i = 1,..,n_2$. (3.1.10)

Se definiamo $\hat{p}_{.j}$ tale che

$\sum_{i=1}^{n_1} \hat{p}_{ij} \cdot r_{ij} = \hat{p}_{.j} \cdot r_{.j}$

$\hat{p}_{.j}$ ha il significato di premio medio da applicare agli assicurati con modalità j per la seconda variabile tariffaria. Sfruttando la (3.1.5), si ottiene la seguente espressione per la condizione di bilanciamento sulle righe: $\hat{p}_{.j} r_{.j} = p_{.j} r_{.j}$ per ogni $j = 1,..,n_2$;

ovvero:

$\hat{p}_{.j} = p_{.j}$ per ogni $j = 1,..,n_2$.

Quindi la condizione di bilanciamento sulle colonne richiede che i premi medi uguaglino gli esborsi medi osservati per gli assicurati aventi la stessa determinazione per la seconda variabile tariffaria. E' interessante

introdurre ancora la definizione di bilanciamento globale. Diremo che il metodo tariffario verifica il bilanciamento globale se:

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\hat{p}_{ij}\cdot r_{ij} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathrm{p}_{ij}\cdot r_{ij} \qquad (3.1.11)$$

Se definiamo $\hat{p}$ tale che:

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\hat{p}_{ij}\cdot r_{ij} = \hat{p}\cdot r$$

$\hat{p}$ ha il significato di premio medio generale. Si ottiene, dunque, la seguente espressione per la condizione di bilanciamento globale:

$$\hat{p}\cdot r = p\cdot r,$$

ovvero

$$\hat{p} = p$$

per cui la condizione richiede che il premio medio generale coincida con l'esborso medio osservato.

E' interessante notare che la condizione (3.1.11) può essere così riscritta:

$$\sum_{j=1}^{n_2}\hat{p}_{.j}\cdot r_{.j} = \sum_{j=1}^{n_2}p_{.j}\cdot r_{.j}$$

e quindi condizione sufficiente affinché il metodo tariffario verifichi il bilanciamento globale è che il metodo verifichi il bilanciamento sulle colonne. Un'analoga condizione può essere ottenuta con riferimento al bilanciamento sulle righe.

Se si adottano come stime delle relatività $\hat{\beta}_{1i}$ con $i = 1,\dots,n_1$ e $\hat{\beta}_{2j}$ con $j = 1,\dots,n_2$ le relatività intuitive (3.1.7) e (3.1.8) non è garantito in generale il bilanciamento sulle singole modalità j. Infatti, con riferimento alla j-esima modalità si ha:

$$\hat{p}_{.j} = \frac{\sum_{i=1}^{n_1}\hat{p}_{ij}\cdot r_{ij}}{r_{.j}} = \frac{\sum_{i=1}^{n_1}\hat{p}\cdot\hat{\beta}_{1i}\cdot\hat{\beta}_{2j}\cdot r_{ij}}{r_{.j}} = \frac{\sum_{i=1}^{n_1}\frac{p_{i.}\cdot p_{.j}}{p}\cdot r_{ij}}{r_{.j}} = p_{.j}\left(\frac{\sum_{i=1}^{n_1}\mathrm{p}_{i.}\cdot r_{ij}}{\mathrm{p}\cdot r_{.j}}\right) = p_{.j}\left(\frac{\sum_{i=1}^{n_1}\mathrm{p}_{i.}\cdot\frac{r_{ij}}{r_{.j}}r}{\mathrm{p}\cdot r}\right) \qquad (3.1.12)$$

Se $\frac{\sum_{i=1}^{n_1}\mathrm{p}_{i.}\cdot\frac{r_{ij}}{r_{.j}}r}{\mathrm{p}\cdot r} \neq 1$ la tariffa non verifica il bilanciamento della modalità j e quindi neppure il bilanciamento globale è garantito. Confrontando l'espressione in parentesi nella (3.1.12) con la (3.1.6) si vede che se la distribuzione degli assicurati nella modalità j $\left(\frac{r_{ij}}{r_{.j}}, i = 1,\dots\dots,n_1\right)$ coincide con la distribuzione marginale degli assicurati sulle classi di rischio $\left(\frac{r_{i.}}{\mathrm{r}}, i = 1,\dots\dots,n_1\right)$ allora il rapporto è uguale a 1. Altrimenti, si intuisce che se nella modalità j si ha una maggiore incidenza di assicurati appartenenti alle classi di rischio peggiori, il rapporto sarà maggiore di 1; altrimenti sarà minore di 1. Nel primo caso il premio medio assegnato ai rischi appartenenti alla j-esima zona sarà maggiore dell'esborso medio osservato nella stessa zona. Nell'ipotesi particolare di indipendenza tra le distribuzioni degli assicurati nelle classi di rischio e nelle zone di circolazione, essendo:

$r_{ij} = \frac{r_{i.}\cdot r_{.j}}{r}$ per ogni $i = 1,\dots\dots,n_1$ e per ogni $j = 1,\dots,n_2$

Si ha:

$$\frac{\sum_{i=1}^{n_1}\mathrm{p}_{i.}\cdot r_{ij}}{\mathrm{p}\cdot r_{.j}} = \frac{\sum_{i=1}^{n_1}\mathrm{p}_{i.}\cdot r_{i.}\cdot r_{.j}}{\mathrm{p}\cdot r\cdot r_{.j}} = \frac{\sum_{i=1}^{n_1}\mathrm{p}_{i.}\cdot r_{i.}}{\mathrm{p}\cdot r} = 1$$

per cui sono soddisfatte le condizioni di bilanciamento per tutte le zone e quindi anche la condizione di bilanciamento globale. Si può verificare in modo analogo che, sotto questa stessa ipotesi, anche le condizioni di bilanciamento per le classi di rischio sono tutte soddisfatte.

3.2. IL METODO DI STIMA DEI TOTALI MARGINALI

L'obiettivo, come già anticipato, è quello di realizzare il bilanciamento per tutti i gruppi di assicurati caratterizzati dalla stessa determinazione per una delle variabili tariffarie, ad esempio per ciascun gruppo di assicurati appartenenti ad una stessa classe di rischio, per ogni gruppo di assicurati appartenenti ad una stessa zona e così via per tutte le classi di rischio e per tutte le zone tariffarie. Si devono quindi trovare le relatività che soddisfano il sistema delle seguenti condizioni

$$\begin{cases} \sum_{j=1}^{n_2} p \cdot \hat{\beta}_{1i} \cdot \hat{\beta}_{2j} \cdot r_{.j} = \sum_{j=1}^{n_2} p \cdot r_{.j} \\ \sum_{i=1}^{n_1} p \cdot \hat{\beta}_{1i} \cdot \hat{\beta}_{2j} \cdot r_{i.} = \sum_{i=1}^{n_1} p \cdot r_{i.} \end{cases} \tag{3.2.1}$$

Questo sistema non ammette soluzioni ma si può, tuttavia, ottenere una soluzione approssimata risolvendo per via iterativa il seguente sistema:

$$\begin{cases} \hat{\beta}_{1i} = \dfrac{p_{i.}/p}{\hat{\beta}_{2j}\frac{r_{ij}}{r_{i.}}} \\ \hat{\beta}_{2j} = \dfrac{p_{j.}/p}{\hat{\beta}_{1i}\frac{r_{ij}}{r_{.j}}} \end{cases} \tag{3.2.2}$$

A partire ad esempio da un insieme di valori $\{\hat{\beta}_{2j} : j = 1, \dots, n_2\}$ scelti arbitrariamente, questi possono venire sostituiti nella prima delle equazioni (3.2.2) in modo da calcolare i valori $\{\hat{\beta}_{1i} : i = 1, \dots, n_1\}$ che le rendono soddisfatte; questi verranno a loro volta sostituiti nella seconda delle (3.2.2) determinando un nuovo insieme di valori per i parametri $\{\hat{\beta}_{2j} : j = 1, \dots, n_2\}$ da inserire nella prima delle (3.2.2), e così via. Diversi insiemi di valori "iniziali" $\{\hat{\beta}_{2j} : j = 1, \dots, n_2\}$ portano a diversi insiemi di soluzioni, in quanto le relatività per il modello moltiplicativo sono determinate a meno di una costante moltiplicativa.

3.3. METODI MULTIVARIATI (GLM).

I modelli lineari generalizzati formano un ampio capitolo della statistica multivariata. L'obiettivo primario di tali modelli è quello di mettere in relazione il valore atteso di un fenomeno aleatorio con un set di osservazioni, dette anche covariate o variabili indipendenti. E' necessario riscontrare e verificare che vi sia tra queste due entità un sufficiente grado di dipendenza. In ambito assicurativo ciò significa che deve esistere un legame tra le variabili di interesse ai fini del calcolo del premio, (numero dei sinistri, importo del danno associato ad un sinistro, ecc.) e le variabili selezionate per personalizzare il premio.

Il sopra citato legame deve essere tradotto in un rigido formalismo matematico, affinché si possa pervenire ad una formula, che riceva in ingresso le caratteristiche del rischio assicurato e restituisca il valore del premio da richiedere al contraente. Detto formalismo matematico, come si evince dal titolo, è di natura lineare, ovvero attraverso una combinazione lineare delle variabili indipendenti si ottiene il valore atteso della variabile aleatoria dipendente, o più precisamente di una sua trasformata.

Proprio in quest'ultima definizione risiede la motivazione a chiamare tali modelli "lineari generalizzati", dunque, in estrema sintesi, si cerca una relazione lineare tra le variabili indipendenti e la variabile risposta espressa in una scala che non è necessariamente quella originaria.

Si indichi con:

- Y_k il k-esimo elemento del vettore aleatorio della variabile risposta $k = 1, \dots, n$
- $X_{k,1}, X_{k,2}, \dots, X_{k,m}$ la k-esima riga della matrice delle m misurazioni (a priori) $k = 1, \dots, n$
- $g: R \to R$ una funzione monotona e invertibile, detta link-function,

allora i modelli lineari generalizzati definiscono la seguente equazione di regressione:

$$g(Y_k) = \beta_0 + \sum_{j=1}^{m} X_{k,j} \cdot \beta_j + \varepsilon_k$$

Dove:

- il vettore $\beta_1, \dots, \beta_m$ rappresenta i parametri da stimare;
- β_0 è l'intercetta del modello ed è comune a tutti i profili di rischio;
- ε_k è invece un termine di errore non catturato dal modello.
- $\beta_0 + \sum_{j=1}^{m} X_{k,j} \cdot \beta_j = \eta_k$ è il previsore lineare, che rappresenta la componente sistematica del modello ed è funzione lineare dei parametri $(\beta_0, \beta_1, \dots\dots, \beta_m)$.

La procedura utilizzata per la stima dei parametri del modello deve essere tale da conseguire come obiettivo quello di minimizzare i residui del modello ε_k, ovvero i valori osservati devono essere abbastanza vicini ai valori teorici forniti dal modello, dunque ottenere un elevato livello di "goodness of fitting".

Una delle ipotesi di base dei modelli lineari generalizzati è che le osservazioni Y_k siano realizzazioni di variabili aleatorie appartenenti alla famiglia esponenziale e dunque, che la funzione di densità possa essere scritta mediante la seguente espressione:

$$f_Y(y;\theta,\varphi) = exp\left\{\frac{(y\theta - b(\theta))}{\varphi/\omega} + c(y,\varphi,\omega)\right\} \quad (3.3.1)$$

Per specifiche funzioni $b(.), c(.)$ si ottengono diverse distribuzioni notevoli, come meglio descritto successivamente.

Il parametro θ ,noto come il parametro canonico, è spesso calcolato secondariamente.
Le suddette distribuzioni sono infatti usualmente parametrizzate rispetto alle seguenti grandezze:

- Il valore atteso della variabile risposta (μ), ottenuto per mezzo dei regressori β;
- Il parametro di dispersione (φ/ω), dove ω è una variabile che esprime il peso delle unità statistiche analizzate, definita nel dominio $(0, +\infty)$.

La stima dei parametri del modello può essere ottenuta massimizzando la funzione di verosimiglianza:

$L(\mu; y; \varphi) = \prod_{k=1}^{n} f_{Y_k}(y_k; \theta_k, \varphi)$,

o, alternativamente, massimizzando la funzione di log-verosimiglianza:

$$l(\mu; y; \varphi) = \sum_{k=1}^{n} log f_{Y_k}(y_k; \theta_k, \varphi) \quad (3.3.2)$$

dove $l(\mu; y; \varphi)$ è una funzione che dipende dalle osservazioni y, dal parametro φ e da μ, dove quest'ultimo rappresenta il vettore dei valori attesi teorici, che sono funzione dei parametri di regressione.

Stimare i parametri massimizzando la funzione di log-verisimiglianza, risponde al requisito di "goodness of fitting", infatti è noto lo stretto legame tra la devianza scalata del modello e la funzione di log-verosimiglianza.

La media e la varianza di Y possono essere facilmente ottenute introducendo la funzione generatrice dei momenti tipica delle famiglie esponenziali (per non appesantire la notazione si limiterà la dimostrazione al caso equipesato):

$m_Y(t; \theta; \varphi; \omega) = exp\left\{\frac{b(\theta + t\varphi) - b(\theta)}{\varphi}\right\}$.

Quindi la distribuzione avrà momenti finiti di ogni ordine e si avrà:

$$E[Y^n] = \frac{d^n}{dt^n} m_Y(t; \theta; \varphi)|_{t=0}$$

$$E[Y] = \mu = \frac{d}{dt} m_Y(t; \theta; \varphi)|_{t=0} = exp\left\{\frac{-b(\theta)}{\varphi}\right\} exp\left\{\frac{b(\theta+t\varphi)}{\varphi}\right\} \frac{b\prime(\theta+t\varphi)}{\varphi} \varphi|_{t=0} = b'(\theta) \quad (3.3.3)$$

$$E[Y^2] = exp\left\{\frac{-b(\theta)}{\varphi}\right\} \frac{d}{dt} [exp\left\{\frac{b(\theta+t\varphi)}{\varphi}\right\} b'(\theta + t\varphi)]|_{t=0}$$

$$= exp\left\{\frac{-b(\theta)}{\varphi}\right\} \left[exp\left\{\frac{b(\theta+t\varphi)}{\varphi}\right\} b'^{(\theta+t\varphi)} b'^{(\theta+t\varphi)} + exp\left\{\frac{b(\theta+t\varphi)}{\varphi}\right\} b''(\theta + t\varphi)\varphi\right] |_{t=0}$$

$$= exp\left\{\frac{-b(\theta)}{\varphi}\right\} \left[exp\left\{\frac{b(\theta)}{\varphi}\right\} b'(\theta) b'(\theta) + exp\left\{\frac{b(\theta)}{\varphi}\right\} b''(\theta)\varphi\right] = (b'(\theta))^2 + b''(\theta)\varphi$$

E dunque:

$$VAR[Y] = E[Y^2] - E[Y]^2 = b''(\theta)\varphi \quad (3.3.4)$$

dove le derivate della funzione $b(.)$ sono intese rispetto a θ.

Assegnata una famiglia esponenziale lineare, si dimostra che la derivata prima della funzione generatrice dei cumulanti (fgc) è monotona crescente, quindi la funzione b' è invertibile.

La funzione di varianza è:

$$V(\mu) = b''\left(b'^{-1}(\mu)\right)$$

La varianza di Y può essere espressa attraverso il parametro μ, per cui:

$$VAR(Y) = \varphi V(\mu)$$

$$V(\mu) = b''(b'^{-1}(b'(\theta))) = b''(\theta) = \frac{VAR(Y)}{\varphi}$$

Si ricava, dunque, che il valore atteso della variabile aleatoria Y, in caso di osservazioni pesate, è fornito dalla seguente espressione:

$$E(Y) = \mu = b'(\theta)$$

ed in maniera analoga, per la varianza si ha:

$$VAR(Y) = b''^{(\theta)} \frac{\varphi}{\omega} = V(\mu) \frac{\varphi}{\omega}$$

Quindi la varianza di Y può essere scritta come il prodotto di due funzioni: la prima, $b''(\theta)$ che dipende solamente dai parametri della distribuzione dai quali dipende anche il valore atteso; la seconda, φ/ω che dipende, invece, dal parametro φ.

Per quanto riguarda il parametro θ, si è detto che il suo nome è parametro canonico: in ogni famiglia esponenziale lineare, la funzione b'^{-1}, trasforma la speranza matematica μ nel parametro canonico θ. Infatti, da $\mu = b'(\theta)$ e dall'invertibilità di b' si ha che

$b'^{-1}(\mu) = \theta$

Scegliendo $g(\mu) = b'^{-1}(\mu)$, come link function si ha:

$\eta_k = g(\mu_k) = \theta_k, k = 1 \dots n$

Tale funzione canonica mette direttamente in collegamento il previsore lineare con il parametro canonico, che è espresso come combinazione lineare delle variabili esplicative.

Inoltre vale:

$g'(\mu) = \frac{1}{b''\left(b'^{-1}(\mu)\right)} = \frac{1}{V(\mu)}$

Nella Tabella 1 per le funzioni canoniche, della densità esponenziale definita nella (3.3.1), per le principali distribuzioni di probabilità.

Tabella 1: Funzioni canoniche al variare della distribuzione di probabilità

	Normale	Poisson	Binomiale	Gamma	Gaussiana Inversa
Notazione	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m,\pi)/m$	$G(\mu,\nu)$	$IG(\mu,\sigma^2)$
Supporto di y	$(-\infty,\infty)$	$0(1)\infty$	$0(1)m/m$	$(0,\infty)$	$(0,\infty)$
Parametro di dispersione: φ	$\varphi = \sigma^2$	1	$1/m$	$\varphi = \nu^{-1}$	$\varphi = \sigma^2$
fgc: b(θ)	$\theta^2/2$	e^{θ}	$\log(1+e^{\theta})$	$-\log(-\theta)$	$-(-2\theta)^{\frac{1}{2}}$
c(y; φ)	$-(y\text{^}2/\varphi + \log(2\pi\varphi))/2$	$-\log y!$	$\log\binom{m}{my}$	$\nu\log(\nu y) - \log y$ $-\log\Gamma(\nu)$	$\frac{1}{2}\left\{\log(2\pi\varphi y^3 + \frac{1}{\varphi y}\right\}$
$\mu(\theta)$ $= E(Y;\theta)$	θ	e^{θ}	$e^{\theta}/(1+e^{\theta})$	$-1/\theta$	$(-2\theta)^{-1/2}$
Legame Canonico	Identità	Log	Logit	Reciproco	$1/\mu^2$
Funzione di Varianza	1	μ	$\mu(1-\mu)$	μ^2	μ^3

3.3.1. STIMA DEI PARAMETRI

Per quanto riguarda la stima dei parametri, tornando al caso generale, la funzione di log-verosimiglianza (3.3.2) può essere riscritta attraverso la seguente:

$l(y;\theta,\varphi) = \frac{1}{\varphi}\sum_k \omega_k\left(y_k\theta_k - b(\theta_k)\right) + \sum_k c(y_k,\varphi,\omega_k)$

Supponiamo, per ora, che il parametro φ sia fissato. Per i modelli per i quali tale parametro non è dato, ciò equivale a imporre una restrizione della log-verosimiglianza, ma ai fini di ottenere la stima di $\boldsymbol{\beta}$ la condizione non è restrittiva. In maniera più precisa, la funzione di verosimiglianza è funzione dei parametri β, piuttosto che dei parametri θ, i quali sono invece collegati al valor medio attraverso la relazione $\mu = b'(\theta)$, che va combinata con la link-function $g(\mu_k) = \eta_k = \sum_j x_{k,j}\beta_j$.

Derivando la funzione l rispetto ai parametri β_j si ottiene:

$$\frac{\delta l}{\delta \beta_j} = \sum_k \frac{\delta l}{\delta \theta_j}\frac{\delta \theta_k}{\delta \beta_j} = \frac{1}{\varphi}\sum_k \omega_k\big(y_k\theta_k - b'(\theta_k)\big)\frac{\delta \theta_k}{\delta \beta_j} = \frac{1}{\varphi}\sum_k \omega_k\big(y_k\theta_k - b'(\theta_k)\big)\frac{\delta \theta_k}{\delta \mu_k}\frac{\delta \mu_k}{\delta \eta_k}\frac{\delta \eta_k}{\delta \beta_j}$$

Dalla sopracitata relazione $\mu = b'(\theta)$ abbiamo che $\frac{\delta \mu_k}{\delta \theta_k} = b''(\theta_k)$. La derivata della relazione inversa è semplicemente il reciproco della derivata, cioè

$$\frac{\delta \theta_k}{\delta \mu_k} = 1/V(\mu_k).$$

Inoltre,

$$\frac{\delta \mu_k}{\delta \eta_k} = \left[\frac{\delta \eta_k}{\delta \mu_k}\right]^{-1} = 1/g'(\mu_k).$$

Combinando tutti i risultati ottenuti, si ha:

$$\frac{\delta l}{\delta \beta_j} = \frac{1}{\varphi}\sum_k \omega_k \frac{y_k - \mu_k}{V(\mu_k)g'(\mu_k)}x_{kj},$$

che è anche detta score-function.

Uguagliando le derivate parziali a zero e moltiplicando per $\boldsymbol{\varphi}$, si ottengono le maximum likelihood equations (ML-equations):

$$\sum_k \omega_k \frac{y_k - \mu_k}{V(\mu_k)g'(\mu_k)}x_{kj} = 0 \quad j = 1,2,\ldots..,m \tag{3.3.5}$$

Ricordando che $\mu_k = g^{-1}\big(\sum_j x_{kj}\beta_j\big)$, il sistema di equazioni (3.3.5) deve essere risolto rispetto alle m incognite β_j .

Per determinare i parametri di distribuzioni diverse dalla Normale, un sistema analitico (come quello dei minimi quadrati utilizzati per la regressione lineare) sarebbe troppo complesso da risolvere, o poco performante in termini di tempi di elaborazione, anche dei calcolatori più evoluti.
Il GLM quindi necessita di un algoritmo iterativo (numerico) (spesso quello di Newton-Rapson), che ha l'obiettivo di massimizzare la funzione di log-verosimiglianza.

Una volta definita la procedura per la stima dei parametri occorre notare che, in analogia con i modelli di regressione lineare, esiste un legame tra il valore atteso condizionato della variabile risposta e il predittore lineare, ed è proprio la funzione link a garantirlo:

$$E(Y|\boldsymbol{x}_k) = g^{-1}(\boldsymbol{x}_k\boldsymbol{\beta}) = g^{-1}(\boldsymbol{\eta}_k) \quad (3.3.6)$$

Dall'equazione (3.3.6) si evince che i modelli lineari generalizzati forniscono una stima del valore atteso di Y condizionato alle caratteristiche dell'individuo definite dal vettore $\boldsymbol{x}_k$.

Si può concludere quindi, riprendendo la notazione del paragrafo 2.5, che i modelli lineari generalizzati definiscono un premio individuale della forma:

$$P_{ij} = P[Y_{ij}] = P[Y|x_{ij}] = E(Y|\boldsymbol{x}_{ij})$$

In un GLM la stima dei regressori β non dipende dal parametro di dispersione φ, ma quest'ultimo è determinato in modo automatico nel processo di stima della massima verosimiglianza.

L'influenza di tali parametri nelle distribuzioni di probabilità appartenenti alla famiglia esponenziale definisce la determinazione della funzione di varianza. La stima del parametro di dispersione, tuttavia, è particolarmente interessante per i nostri fini, in quanto fa sì che il GLM non si limiti a fornire una stima del solo valore atteso condizionato alle caratteristiche dell'individuo, ma anche dell'intera distribuzione di probabilità condizionata. Si fa notare che nel caso di Poisson, la sola stima del valore atteso condizionato fosse già sufficiente per la stima della distribuzione di probabilità condizionata di Y, poiché nel caso di distribuzioni di probabilità mono-parametriche la media è una statistica sufficiente per la conoscenza dell'intera distribuzione.

3.4. LA PERSONALIZZAZIONE DEL PREMIO CON I GLM.

3.4.1. LA FATTORIZZAZIONE DELLA QUOTA DANNI

I GLM forniscono una stima del valore atteso condizionato della variabile risposta. Per applicarli alla tariffazione, l'idea più semplice è quella di definire una distribuzione di probabilità coerente con le caratteristiche di Y e ottenere direttamente una stima di $E[Y|\boldsymbol{x}_k]$. Tale metodologia è poco utilizzata però nella pratica, poichè la natura semicontinua della variabile Y, richiederebbe un'ipotesi di appartenenza alla famiglia tweedie, che non è di semplice trattabilità.

La metodologia più utilizzata, sfrutta la proprietà di fattorizzazione del valore atteso della quota danni. Formalmente, considerando le variabili casuali:

- $N = numero\ di\ sinistri$;
- $Z_i = importo\ del\ danno\ associato\ ad\ un\ sinistro$;

sotto le seguenti ipotesi:

- Z_i indipendente da N;
- Z_i sono indipendenti e identicamente distribuite: $F_{Z_i} = F_Z$;

è possibile dimostrare che:

$E[Y] = E[N] \cdot E[Z]$

(3.4.1)

Si effettua, quindi, un GLM per la frequenza sinistri (nella pratica si fa un ipotesi Poisson, per la distribuzione di N) e un GLM per il costo medio (nella pratica si fa un ipotesi Gamma, per la distribuzione di Z).

Le ipotesi sui due modelli sono le seguenti:

- Modello di frequenza
  - Ipotesi distributiva: Poisson
  - Legame canonico: Logaritmo
  - Funzione link scelta: Logaritmo
- Modello di costo medio:
  - Ipotesi distributiva: Gamma
  - Legame canonico: Reciproca
  - Funzione link scelta: Logaritmo

La funzione link adottata usualmente per fini tariffari è infatti quella logaritmica. Per le proprietà della funzione logaritmo, la struttura dei coefficienti (o regressori), che saranno creati dal GLM in output sarà moltiplicativa.

3.4.2. IL PROBLEMA DEI SINISTRI LARGE

L'approccio descritto al paragrafo precedente è ulteriormente modificato, in quanto occorre tenere conto del problema dei sinistri large, (sinistri di importo particolarmente elevato) che con la loro presenza compromettono, spesso in modo decisivo, la bontà di adattamento del modello di costo medio. Includere tali sinistri nella base dati per il calcolo dei premi, significa attribuire ad alcune tipologie di contratti una sinistrosità che non gli appartiene, come sarà ampiamente dimostrato nell'applicazione. Risulta quindi evidente la necessità di dividere i dati in due gruppi da modellare e trattare in maniera differente. La procedura è la seguente:

- Individuazione di una soglia K per ripartire i sinistri in due gruppi:
  a. Sinistri Attritional: sinistri di importo inferiore ad una soglia K prefissata;
  b. Sinistri Large: Sinistri oltre una soglia K prefissata.
- Definizione di una decomposizione del valore atteso della quota danni più articolata rispetto a quella espressa nella (3.4.1), che permetta di trattare in maniera più efficace il problema dei sinistri large:

$$E[Y] = E[N] \cdot E[Z] = E[N] \cdot \{E[Z|Z \leq K] \cdot P[Z \leq K] + E[Z|Z > K] \cdot P[Z > K]\} \quad (3.4.2)$$

L'idea è quella di eseguire tanti modelli GLM, quante sono le grandezze introdotte nella (3.4.2), con le seguenti ipotesi:

- Modello per la stima della frequenza attesa $E[N]$

    - Ipotesi distributiva: Poisson
    - Legame canonico: Logaritmo
    - Funzione link scelta: Logaritmo
- Modello per la stima del costo medio attritional $E[Z|Z \leq K]$:
    - Ipotesi distributiva: Gamma
    - Legame canonico: Reciproca
    - Funzione link scelta: Logaritmo
- Modello per la stima di $P[Z \leq K]$:
    - Ipotesi distributiva: Binomiale
    - Legame canonico: Logit
    - Funzione link scelta: Logit

Per quanto riguarda la stima di $E[Z|Z > K]$ si ritiene, in genere, che i sinistri large non dipendano dalle caratteristiche del rischio assicurato, ma dalle circostanze nelle quali avviene il sinistro. È usuale, dunque, stimare tale componente uguale per tutti gli assicurati, in base ai dati osservati.

Nella pratica si stima una soglia K sfruttando la teoria dei valori estremi utilizzando tecniche che esulano da questo lavoro di tesi. Una volta fissata la soglia si costruisce il risarcimento medio per profilo di rischio a partire dall'equazione (3.4.2), il che equivale ad eseguire tanti GLM, quanti sono gli elementi costituenti l'equazione stessa. Tuttavia per ragioni di semplicità, e sulla base della considerazione che i sinistri punta non si ipotizzano dipendenti dalle caratteristiche dell'assicurato, il costo medio dei sinistri large si stima a partire dai dati osservati indipendentemente dalle caratteristiche dell'individuo.

4. QUANTILE REGRESSION

4.1. COME E' NATA LA QUANTILE REGRESSION

Nelle procedure statistiche parametriche, per indagare un fenomeno si cerca di descrivere il processo che genera i dati attraverso un elemento di una classe di modelli parametrici. Il valutatore cerca statistiche, cioè funzioni dei dati campionari, per ottenere stimatori la cui distribuzione sarà concentrata il più possibile intorno al vero valore del parametro. Questa procedura sottintende che il modello sottostante sia corretto; se, come quasi sempre accade, il modello parametrico non sintetizza efficacemente il fenomeno, l'obiettivo dovrebbe essere quello di usare stimatori la cui distribuzione sia poco sensibile ad osservazioni "anomale", ovvero stimatori "*robusti*". Nel linguaggio statistico il termine "*robustezza*" definisce una sorta di "resistenza" delle procedure statistiche a scostamenti dalle ipotesi iniziali del modello. In alcuni casi fondamentali, la "robustezza" non è una caratteristica degli stimatori di uso comune. Consideriamo il noto problema di dover stimare un vettore di parametri incogniti β da un campione di osservazioni indipendenti su variabili aleatorie $Y_1, Y_2, \dots \dots, Y_n$, la cui funzione di distribuzione è F, e $\boldsymbol{x}_k \;\; k = 1, \dots \dots, n$ denota le righe di una matrice disegno $(n \; x \; m)$.

Se F fosse precisamente nota, è possibile mostrare che lo stimatore di massima verosimiglianza è efficiente nel senso di Cramer-Rao. In particolare, quando F è una Gaussiana, Rao ha mostrato che lo stimatore dei minimi quadrati $\dot{\beta}$ è quello di varianza minima nella classe degli stimatori non distorti. Sfortunatamente, l'estrema sensibilità dello stimatore dei minimi quadrati, anche ad una quantità modesta di "outliers", lo rende uno stimatore poco robusto in molte situazioni a coda lunga.

L'idea è quella d'introdurre una nuova classe di stimatori robusti, rispetto a quello dei minimi quadrati, per il modello lineare, che siano comparabili in termini di efficienza a $\dot{\beta}$ nel modello Gaussiano e che siano più performanti su un ampia classe di modelli non Gaussiani.

Il bisogno di alternative robuste alla media semplice (statistica di fatto definita dallo stimatore dei minimi quadrati) è apparsa fin dal XVIII secolo. La mediana, altre medie troncate e più complicate combinazioni lineari di statistiche ordinali, erano comunemente usate nei calcoli astronomici. Dal 1921, Gauss ha mostrato che la media semplice fornisce "la più probabile" stima del location parameter da un campione casuale con densità di probabilità proporzionale a $e^{-x^2/2\sigma^2}$, ma questo risultato era esplicitamente una razionalizzazione ex post dell'uso della media semplice, piuttosto che una prova della validità empirica di questa particolare legge dell'errore. Infatti fu notato da un buon numero di autori, che esisteva un numero abbastanza cospicuo di fenomeni con distribuzioni degli errori con code più lunghe, di quella della distribuzione Gaussiana. In tali casi apparve auspicabile scegliere stimatori che fossero meno influenzati da osservazioni estreme.

Molte figure illustri (Gauss, Laplace e Legendre) suggerirono che la minimizzazione degli scarti in valore assoluto potesse essere preferibile a quella degli scarti al quadrato, quando alcune osservazioni campionarie fossero difficilmente riconducibili all'ipotesi gaussiana.

Il punto è che possono esistere stimatori non lineari, o distorti, migliori rispetto a quello dei minimi quadrati, per modelli non Gaussiani. Per esempio, la media pesata dei quantili al livello di probabilità 1/3, 1/2 e 2/3 con pesi rispettivamente pari a 0,3, 0,4 e 0,3 ha un'efficienza asintotica di quasi l'80% per distribuzioni Gaussiane, Laplace, logistica e Di Cauchy (l'efficienza asintotica della media aritmetica nel caso Gaussiano costituisce l'unità). Al contrario, la media semplice ha un'efficienza asintotica pari ad 1 nel caso Gaussiano, ma è efficiente la metà della mediana nel caso di distribuzione Laplace e ha efficienza 0 per la distribuzione di Cauchy. Dunque, questi stimatori introdotti sono statistiche inefficienti in alcuni modelli parametrici, ma in pratica, possono essere preferibili a stimatori ritenuti convenzionalmente "ottimali", come la media semplice, se c'è incertezza nella forma di distribuzione che genera il campione.

Prima di introdurre la Quantile Regression si fa notare che il primo tentativo nella storia di effettuare una regressione è molto vicino al concetto di Quantile Regression. Infatti il modello dei minimi quadrati è datato in corrispondenza al lavoro pubblicato da Legendre nel 1805, mentre quello di Boscovich, a cui si farà riferimento nel seguito, è datato circa mezzo secolo prima.

Il problema su cui si era cimentato Boscovich era quello dell'ellitticità della terra. Per dimostrare la validità della tesi, l'autore aveva fatto riferimento alle misure che compaiono in Figura 7. Ognuna di esse rappresenta la lunghezza dell'arco di meridiano corrispondente a un grado di latitudine in 5 città: da Quito (equatore) a Lapland. Risultava chiaro da queste misurazioni che la lunghezza dell'arco, cresceva muovendosi dall'equatore al polo, confermando la congettura. Tuttavia non era chiaro come le cinque misure dovessero essere combinate per stimare l'ellitticità della terra.

Figura 7: dati esperimento Boscovich

Location	latitude	$\sin^2$ (latitude)	arc-length
Quito	$0^\circ\ 0'$	0	56751
Cape of Good Hope	$33^\circ\ 18'$	0.2987	57037
Rome	$42^\circ\ 59'$	0.4648	56979
Paris	$49^\circ\ 23'$	0.5762	57074
Lapland	$66^\circ\ 19'$	0.8386	57422

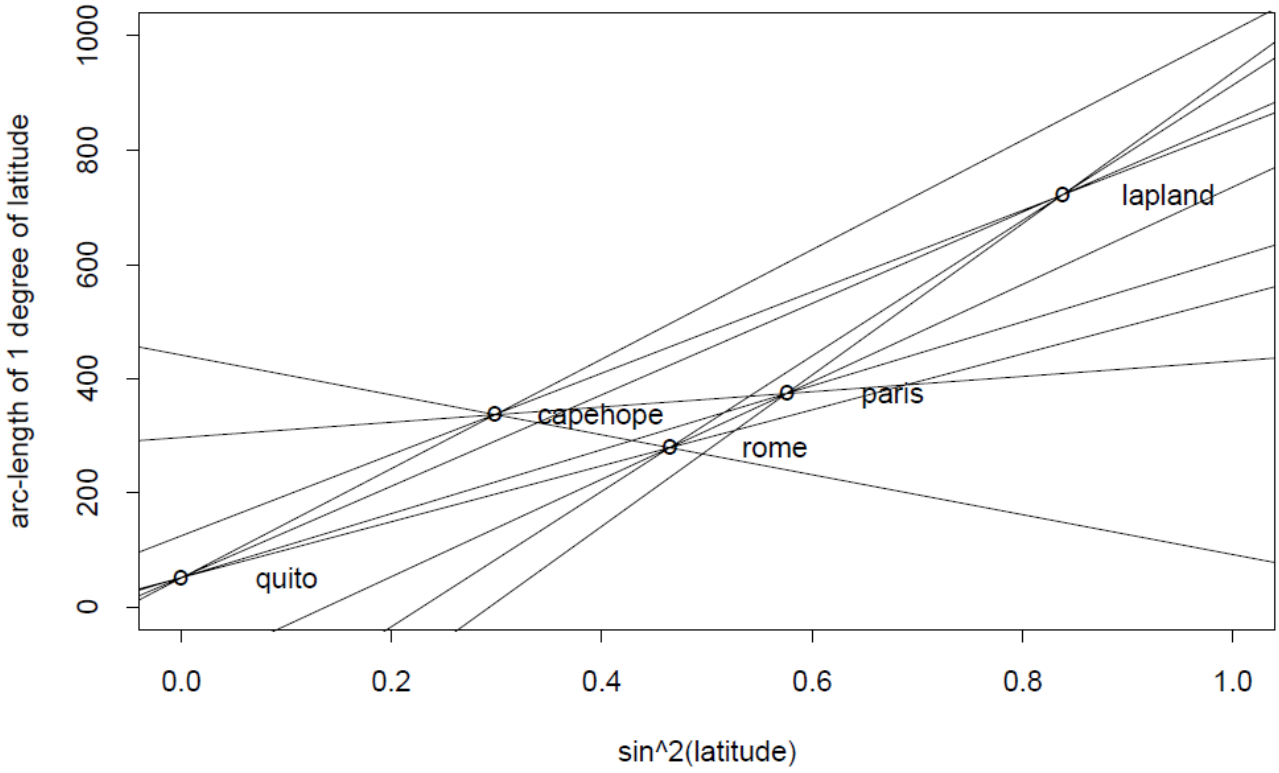
Si dimostra che per archi piccoli vale l'approssimazione:

$$y = a + b \cdot sin^2\lambda,$$

dove y è la lunghezza dell'arco e λ è la corrispondente latitudine. Il parametro a può essere interpretato come la lunghezza dell'arco corrispondente a un grado di latitudine all'equatore e b come l'eccedenza della lunghezza di un grado di latitudine al polo, rispetto a quella all'equatore. L'ellitticità può essere stimata come $\eta = \frac{3a}{b}$. Boscovich decise di utilizzare le osservazioni a due a due per ottenere una stima di a e b, calcolando, quindi, le 10 stime possibili (ovvero le 10 possibili equazioni di una retta passanti per due punti, tali soluzioni di seguito vengono definite **pairwise slopes**). In Figura 8 sono rappresentate le 10 possibili rette.

Si vede (Figura 8) che alcune di queste rette sembrano assolutamente poco plausibili, specialmente quella passante per Città del Capo e Roma. Boscovich decise di riportare una stima finale basata su una media dei 10 risultati possibili per b, mentre la stima di a fu presa direttamente pari alla lunghezza dell'arco in corrispondenza di Quito (ovvero in corrispondenza dell'equatore). La stima definiva un ellitticità pari rispettivamente a $1/155$.

Figura 8: Rette di regressione stimate, considerando i cinque punti presi a due a due



È curioso osservare che lo stimatore dei minimi quadrati per (a, b) può anche essere espresso come media ponderata delle pairwise slopes.

Si indicizzino le 10 coppie con h e si definisca:

$b(h) = X(h)^{-1} y(h)$

dove per il nostro semplice modello bivariato ed $h = (i, j)$

- $X(h) = \begin{pmatrix} 1 & x_i \\ 1 & x_j \end{pmatrix}$
- $y(h) = \begin{pmatrix} y_i \\ y_j \end{pmatrix}$

Allora si può scrivere lo stimatore dei minimi quadrati $(\hat{a}, \hat{b})$ nel seguente modo:

$(\hat{a}, \hat{b}) = \sum_h \omega(h)\, b(h)$

Dove

$\omega(h) = \frac{|X(h)|^2}{\sum_h |X(h)|^2}$

Il secondo tentativo di Boscovich, arrivò due anni dopo e fu molto vicino alla Quantile Regression. Infatti egli suggerì di ottenere le stime di a e b, minimizzando la somma degli scarti assoluti, con il vincolo che la somma degli scarti fosse nulla. Il vincolo richiedeva che la retta passasse per il centroide delle osservazioni $(\bar{x}, \bar{y})$, ossia per le medie di x e y. Avendo ridotto il problema a una regressione passante per l'origine, attraverso il vincolo, si può immaginare di arrivare alla soluzione, ruotando una retta passante per la nuova origine $(\bar{x}, \bar{y})$, fintanto che sia minima la somma degli scarti assoluti. In sintesi il problema è ridotto alla stima del solo coefficiente angolare

Questo può essere visto algebricamente come il calcolo della mediana pesata. Per ogni punto possiamo calcolare:

$b_k = \frac{y_k - \bar{y}}{x_k - \bar{x}}$

e associare a ciascun coefficiente angolare il peso $\omega_k = |x_k - \bar{x}|$. Ora, siano $b_{(k)}$ l'insieme delle stime ordinate e $\omega_{(k)}$ i rispettivi pesi; si calcola il più piccolo j, detto j^*, tale che:

$\sum_{k=1}^{j} \omega_{(k)} > \frac{1}{2} \sum_{k=1}^{n} \omega_{(k)}$.

Lo stimatore di Boscovich $b_{(j^*)}$ definisce il cosidetto "metodo di situazione" ed è un ibrido tra media e mediana; infatti l'intercetta è stimata con una media e il coefficiente angolare come una mediana.

Il lavoro di Edgeworth nel 1888 eliminò il vincolo di Boscovich sull'intercetta e propose di minimizzare la somma degli scarti assoluti, sia attraverso l'intercetta che attraverso il coefficiente angolare. Tale metodo fu denominato della "doppia mediana", affermando che si sarebbe potuto estendere a "plural median method".

Quello che Edgeworth intendeva dire con "plural median method" era che, una volta definita una regressione per la mediana, dovevano esistere analoghe regressioni per tutti i quantili.

4.2. LA DEFINIZIONE DEL QUANTILE COME SOLUZIONE DI UN PROBALEMA DI MINIMO.

Per comprendere meglio la ratio sottostante la Quantile Regression, il punto di partenza è una definizione elementare di quantile campionario che, dato un campione ordinato, può essere prontamente esteso al modello lineare.

Come sopra, sia $\{y_k: \ k = 1, \ldots\ldots, n\}$ un campione casuale proveniente da una variabile aleatoria Y, con funzione di distribuzione F.

Il quantile di livello θ $(0 < \theta < 1)$, può essere definito come soluzione del problema di minimizzazione:

$$min_{a \in R}\{\psi(a)\} = min_{a \in R}\left\{\sum_{k:y_k \geq a} \theta \, |y_k - a| + \sum_{k:y_k < a} (1-\theta) \, |y_k - a|\right\} \qquad (4.2.1)$$

E' possibile mostrare il tutto, nel caso in cui la variabile oggetto di studio sia assolutamente continua: in tale frangente, infatti, la funzione di perdita (4.2.1) diventa:

$$h(a) = \theta \int_a^{\infty} (y-a) dF_Y(y) - (1-\theta) \int_{-\infty}^{a} (y-a) dF_Y(y)$$

Per la condizione del primo ordine, la funzione di perdita viene minimizzata ponendo la derivata prima uguale a zero; ovvero:

$$\frac{dh(a)}{da} = \frac{\theta \int_a^{\infty} (y-a) dF_Y(y)}{da} - \frac{(1-\theta) \int_{-\infty}^{a} (y-a) dF_Y(y)}{da}$$

$$= -\theta \int_a^{\infty} dF_Y(y) + (1-\theta) \int_{-\infty}^{a} dF_Y(y)$$

$$= -\theta\big(1 - F_Y(a)\big) + (1-\theta) F_Y(a)$$

$$= F_Y(a) - \theta = 0$$

Essendo la derivata seconda maggiore di zero, la funzione di perdita attesa è convessa e viene minimizzata se e solo se $F_Y(a) = \theta$, ovvero se $F_Y^{-1}(\theta) = a$

Una volta definito il quantile come soluzione del problema di minimizzazione (4.2.1), si introduce una definizione del problema stesso, in forma compatta, attraverso l'introduzione di una *check function:*

$$\rho_\theta(u) = u\big\{\theta - I\{u < 0\}\big\}, \tag{4.2.2}$$

dove I è la variabile indicatrice.

Per cui è possibile riscrivere il problema (4.2.1) come:

$$min_{a \in R}\{\psi(a)\} = min_{a \in R}\{\sum_{k=1}^{n} \rho_\theta(y_k - a)\} \tag{4.2.3}$$

È noto che anche la media aritmetica è definibile come soluzione di un problema di minimo, infatti essa è quella statistica che minimizza la somma degli scarti al quadrato.

In accordo con Koenker (2005) e Hao-Naiman (2007), il problema di minimizzazione della funzione (4.2.3), può essere trasformato in un problema di programmazione lineare, introducendo $2n$ variabili artificiali $\{u_k^+, u_k^-: 1, \dots, n\}$, che rappresentino le parti positive e negative del vettore dei residui, ovvero:

$$u_k^+ = (y_k - a)^+ = \begin{cases} y_k - a, & y_k \geq a \\ 0 & altrimenti \end{cases}$$

$$u_k^- = (a - y_k)^+ = \begin{cases} a - y_k, & y_k < a \\ 0 & altrimenti \end{cases}$$

Il problema di ottimizzazione (4.2.3) potrà essere definito nel seguente modo:

$$\min \theta \cdot \mathbf{1} \cdot \boldsymbol{u}^+ + (1-\theta) \cdot \mathbf{1} \cdot \boldsymbol{u}^- \tag{4.2.4}$$

$$\begin{cases} \mathbf{1}a + \boldsymbol{u}^{+} - \boldsymbol{u}^{-} = \boldsymbol{y} \\ \boldsymbol{u}^{+} \geq 0 \\ \boldsymbol{u}^{-} \geq 0 \end{cases}$$

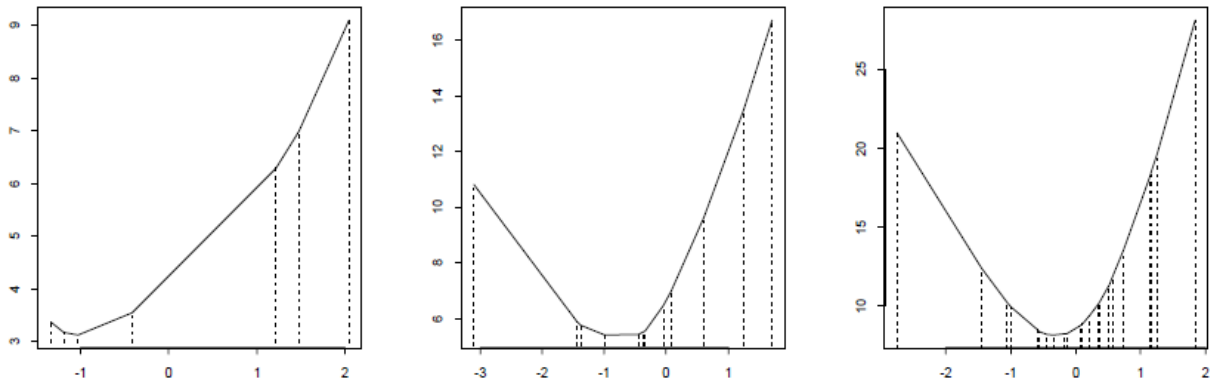
dove $\mathbf{1}$, rappresenta un vettore unitario n-dimensionale

Dai grafici in Figura 9 è possibile vedere che il punto di minimo, se è unico, si trova in corrispondenza di una singola osservazione y_k. In Figura 9, infatti, è rappresentato il problema di ottimizzazione che definisce il quantile a livello $\theta = 1/3$, dove le y_k costituiscono un campione estratto da una normale standard, di numerosità rispettivamente pari a 7,12, e 23 (le linee verticali tratteggiate rappresentano le osservazioni del campione).

Si può notare immediatamente che, poiché 12 è divisibile per 3, la funzione obiettivo è piatta in corrispondenza del suo minimo nella seconda figura. Si conclude quindi che, in questo caso, si ha un intervallo di soluzioni tra la quarta e la quinta realizzazione di Y. Si può osservare che, in ogni caso, il modello presenterà almeno un residuo nullo.

Il grafico della funzione obiettivo è convesso e lineare a tratti con vertici sui valori osservati y_k. Quando a passa attraverso una delle y_k, la pendenza della funzione obiettivo cambia esattamente di 1, poiché un contributo di $\theta - 1$ è sostituito da θ o viceversa.

Figura 9: Funzione obiettivo calcolata per $\theta = 1/3$ su campioni di numerosità diversa



Data la non derivabilità della funzione obiettivo proprio nei punti di minimo globale, risulta impossibile applicare l'annullamento della derivata prima come condizione per la definizione di tali punti. Tale criticità è superata dall'introduzione delle derivate direzionali di ψ.

L'ottimalità si ha se la derivata sinistra e destra:

$$\psi'(a,1) = \lim_{h \to 0} \big(\psi(a+h) - \psi(a)\big)/h = \sum_{k=1}^{n} (I(y_k < a + h) - \theta)$$

ed

$$\psi'(a,-1) = \lim_{h\to 0}\big(\psi(a-h)-\psi(a)\big)/h = \sum_{k=1}^{n}\big(\theta - I(y_k < a-h)\big)$$

siano entrambi non negative, il che significa che la funzione cresce qualunque sia la direzione in cui ci si muova.

La condizione affinchè le due derivate siano entrambe non negative è che $n\theta$ cada nell'intervallo chiuso $[N^-, N^+]$, dove:

$N^{\pm} = \#\{y_k < a \pm \varepsilon\}$, con ε piccolo a piacere.

- Quando $n\theta$ non è intero c'è un unico valore di a che soddisfa questa condizione.
- Quando la variabile di partenza è assolutamente continua, questa soluzione unica corrisponde ad un unico elemento del campione ordinato; se la variabile non è continua la soluzione è sempre unica, ma potrebbero esserci più y_k uguali a quel valore. In questo caso il modello avrà un numero di residui pari a zero maggiore del numero di parametri da stimare.

Una volta definito il quantile campionario, come soluzione del problema di minimizzazione (4.2.4) è interessante considerare il suo problema duale, in quanto la soluzione di quest'ultimo sarà fondamentale per la costruzioni di intervalli di confidenza dei parametri della Quantile Regression. Il problema duale al (4.2.4) ha come soluzione, la funzione generatrice dei ranghi delle osservazioni introdotta da Hajek e Sidak (1967).

Si definisce dunque il problema duale:

$$\max \boldsymbol{y}a \qquad (4.2.5)$$

$$\begin{cases} \mathbf{1}a = (1-\theta)n \\ a \in [0,1]^n \end{cases}$$

Esattamente come il problema primale può essere visto come l'ottimizzazione che genera il quantile campionario, il problema duale è l'ottimizzazione che genera il rango delle osservazioni.

Risulta chiaro dalla (4.2.5) che per $\theta = 0$ sarà che tutte le $\hat{a}_k(0) = 1$ e similmente per $\theta = 1$ tutte le

$\hat{a}_k(1) = 0$.

Partendo da $\theta = 0$ al crescere di θ, le $\hat{a}_k$ diminuiscono secondo il seguente schema: inizialmente si pone l'attenzione su $y_{(1)} = min\{y_1, \dots\dots, y_n\}$, dato che facendo decrescere il suo peso, questo avrà meno impatto sulla somma $\boldsymbol{y}a$; quindi se $y_{(1)} = y_j$, allora mentre θ cresce, a_j deve decrescere per soddisfare il vincolo. a_j continua a diminuire fintanto che $\theta = 1/n$, poiché a quel punto a_j ha raggiunto lo zero e sotto quel valore non può scendere a causa del vincolo . Una volta che $\theta = 1/n$ se esso continua a crescere, si considererà l'osservazione $y_{(2)}$, ovvero la seconda realizzazione più piccola e il suo peso verrà mano a mano abbassato nel medesimo modo. Questo processo per $\theta = 1$, continua fintanto che tutte le osservazioni abbiano peso nullo.

La funzione $\hat{a}_k(\theta)$, soluzione del problema duale è detta rank score function, e prende la forma:

$$\hat{a}_k(\theta) = \begin{cases} 1 & if\ \theta \leq (R_k - 1)/n \\ R_k - \theta n & if (R_k - 1)/n < \theta \leq R_k/n \\ 0 & if\ \theta > R_k/n \end{cases} \quad (4.2.6)$$

dove R_k è il rango di $y_{(k)}$.

Questa funzione coincide con la funzione generatrice dei ranghi introdotta da Hajek e Sidak (1967). Come vedremo nei paragrafi successivi, essi hanno costruito un approccio naturale per la definizione dei ranghi e di test statistici basati sui ranghi.

4.3. LA QUANTILE REGRESSION COME STIMA DEI QUANTILI CONDIZIONATI

Le osservazioni che si sono sviluppate finora sostengono che il quantile possa essere espresso come soluzione di un semplice problema di ottimizzazione; pertanto l'obiettivo successivo è quello di definire modelli più generali di stima di quantili condizionati. I minimi quadrati offrono un buon punto di partenza per lo sviluppo successivo.

Introducendo la notazione:

- $\boldsymbol{x}_k = (x_{k1}, \ldots\ldots\ldots, x_{km})$ un vettore riga di variabili indipendenti $(k = 1, \ldots\ldots, n)$, costituenti una "design matrix" X nota,
- β il set di parametri dell'equazione di regressione ai minimi quadrati,
- β_θ il set di parametri dell'equazione di regressione quantilica,
- $\{y_k:\ k = 1, \ldots\ldots, n\}$ un campione casuale, generato dalla funzione di distribuzione F.

Sapendo che la media aritmetica risolve il problema di minimo:

$min_{\mu \in R} \sum_{k=1}^{n} (y_k - \mu)^2$,

se si è interessati ad esprimere la media condizionata di Y rispetto a X, come:

$\mu(\boldsymbol{x}) = E[Y|\boldsymbol{x}] = \boldsymbol{x}'\boldsymbol{\beta}$,

allora si può stimare β risolvendo:

$min_{\beta \in R^m} \sum_{k=1}^{n} (y_k - \boldsymbol{x}_k'\beta)^2$.

Similmente, dato che il quantile campionario a livello θ è la statistica che risolve il problema (4.2.3), si introduce la funzione quantile condizionato:

$Quant_\theta(Y|\boldsymbol{x}) = \boldsymbol{x}'\beta_\theta$ (ssss)

Per la stima di β_θ occorrerà risolvere:

$min_{\beta_\theta \in R^m}\big(\psi(\beta_\theta)\big) = min_{\beta_\theta \in R^m} \sum_{k=1}^{n} \rho_\theta(y_k - \boldsymbol{x}_k'\beta_\theta) \leftrightarrow$

$min_{\beta_\theta \in R^m}\{\sum_{k:y_k \geq \boldsymbol{x}_k\beta_\theta} \theta\,|y_k - \boldsymbol{x}_k'\beta_\theta| + \sum_{k:y_k < \boldsymbol{x}_k\beta_\theta}(1-\theta)\,|y_k - \boldsymbol{x}_k'\beta_\theta|\}.$ (4.3.1)

In tal senso la Quantile Regression fornisce in output una stima di $Quant_\theta(Y|\boldsymbol{x})$, ovvero una stima del quantile condizionato a livello θ della variabile risposta

Questo problema di minimo è il nucleo dell'argomentazione definita da Koenker e Bassett (1978).

A questo punto la Quantile Regression può essere riformulata come un problema di programmazione lineare

$$\min \theta \cdot \mathbf{1} \cdot \boldsymbol{u}^+ + (1-\theta) \cdot \mathbf{1} \cdot \boldsymbol{u}^- \quad (4.3.2)$$

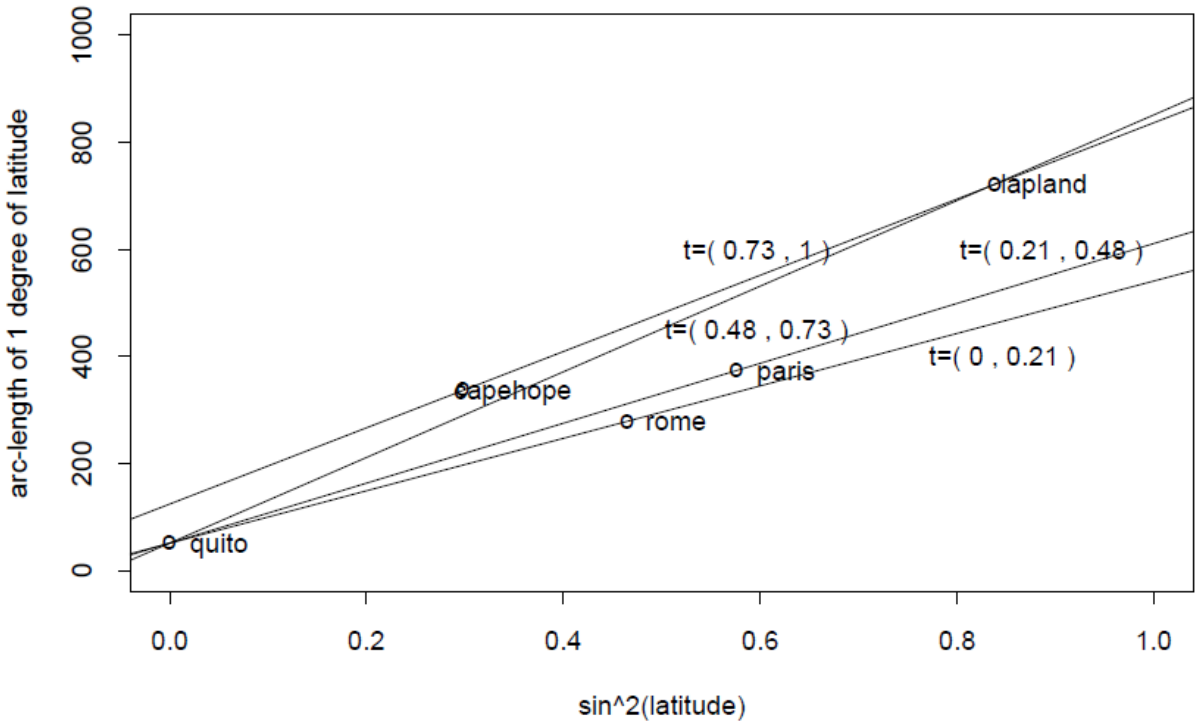
$$\begin{cases} \mathbf{x}'\beta_\theta + \boldsymbol{u}^+ - \boldsymbol{u}^- = \boldsymbol{y} \\ \boldsymbol{u}^+ \geq 0 \\ \boldsymbol{u}^- \geq 0 \end{cases}$$

Si nota dunque che il problema (4.2.4) è un caso particolare del problema (4.3.2) in cui il modello di regressione è un modello con sola intercetta (di seguito "**location model**")

Per comprendere alcune importanti proprietà delle soluzioni β_θ^* del problema (4.3.2), che da ora verranno definiti "**regression quantiles**", si illustrano queste ultime nel caso dei dati di Boscovich (Figura 7). In Figura 10 si rappresentano tutte le soluzioni regression quantiles per questi dati. Delle 10 rette passanti per i punti in Figura 8, la Quantile Regression ne seleziona solo 4. Risolvendo la (4.3.2) per $\theta \in (0; 0{,}21)$ c'è un'unica soluzione, ovvero la retta passante per Quito e Roma. In corrispondenza di $\theta = 0{,}21$, la soluzione si sposta, e nell'intervallo $(0{,}21; 0{,}48)$ si ha la soluzione caratterizzata dalla retta passante per Quito e Parigi.

Si nota, che se nel caso della Quantile Regression sul location model quando la soluzione è unica si ha almeno un residuo nullo, in questo secondo caso in cui si stimano due parametri, almeno due residui saranno sicuramente nulli, in quanto la retta soluzione passa sempre per due dei punti osservati.

Figura 10: Risultati della Quantile Regression sull'esperimento di Boscovich



Nel caso dei quantili campionari ordinari, che sono equamente distribuiti nell'intervallo $[0,1]$, ogni distinta statistica ordinale, occupa un intervallo di lunghezza pari a $1/n$. Nel caso della Quantile Regression le lunghezze degli intervalli sono irregolari e dipendono dalla matrice X e dai valori assunti da y.

Le coppie di punti, definiscono statistiche ordinali e servono a definire la stima della funzione lineare quantile condizionato; inoltre nella terminologia della programmazione lineare, queste soluzioni sono di base e corrispondono ai vertici del poliedro generato dall'insieme dei vincoli.

Se si immagina il piano che rappresenta la funzione obiettivo (4.3.2), mentre ruota al variare di θ, possiamo visualizzare la soluzione , mentre passa da un vertice del poliedro all'altro ed è chiaro quindi che ogni vertice rappresenti una coppia delle osservazioni campionarie.

Inoltre è possibile notare come la Quantile Regression preservi le caratteristiche di robustezza tipiche del quantile ordinario: se perturbassimo le statistiche ordinali sopra (o sotto) la mediana, in modo tale che esse restino sopra (o sotto) la mediana, la posizione della mediana non cambierebbe. Ovvero se modificassimo la latitudine di Lapland in aumento, la mediana risulterebbe invariata.

Risulta chiaro che molte delle intuizioni sulla regressione dei quantili prendono forma dall'interpretazione geometrica del metodo dei minimi quadrati. Sostituendo la somma degli scarti al quadrato con la somma degli scarti assoluti, si introduce un nuovo problema, ma molte utili caratteristiche persistono. Sulla scorta di quanto detto relativamente al location model, si definisce la procedura attraverso la quale si perviene alla stima dei parametri, soluzioni del problema (4.3.2).

Si consideri la derivata direzionale della funzione obiettivo $\psi(\beta_\theta)$:

$\nabla\psi(\beta_\theta, \omega) = \frac{d}{dt}\ \psi(\beta_\theta, t\omega)|_{t=0}$

$= \frac{d}{dt}\sum_{k=1}^{n} u_k(\beta_\theta, t\omega)[\theta - I\{u_k(\beta_\theta, t\omega) < 0\}]\ |_{t=0}$

$= -\sum_{k=1}^{n} \varphi^*\ (y_k - \boldsymbol{x}_k'\beta_\theta, -\boldsymbol{x}_k'\omega)\boldsymbol{x}_k'\omega,$

dove

$$\varphi^*(u; z) = \begin{cases} \theta - I\{u < 0\} & u \neq 0 \\ \theta - I\{z < 0\} & u = 0 \end{cases}$$

Se $\nabla\psi(\beta_\theta, \omega) \geq 0$ per ogni direzione $\omega \in R^m$, con $\|\omega\| = 1$, allora β_θ^* minimizza $\psi(\beta_\theta, \omega)$.

Si tratta di una perfetta generalizzazione della condizione per funzioni smooth $\nabla\psi(\beta_\theta) = 0$, che si deve introdurre a causa della non differenziabilità di ψ.

Un importante caratteristica di β_θ^*, come ormai sarà stato intuito, è che se si stimano m parametri, almeno m residui saranno nulli.

Nella terminologia della programmazione lineare, questo sottoinsieme di m elementi si chiama soluzione di base, geometricamente esse corrispondono ai vertici del poliedro generato dall'insieme dei vincoli. Risulta quindi chiaro, che la soluzione sarà unica quando l'iper-piano che rappresenta la funzione obiettivo, tocca solo un vertice del poliedro e sarà multipla quando l'iper-piano stesso si poggerà su un bordo o addirittura su un intera facciata del poliedro. Si può affermare che quando la soluzione non è unica, le soluzioni di base assumono un ruolo fondamentale, poiché ogni elemento dell'insieme soluzione $B^*(\theta)$, potrà essere ottenuto come combinazione lineare di soluzioni di questo tipo. Koenker e Basset (1978) hanno definito una forma chiusa per descrivere la condizione di unicità della soluzione in un problema del tipo (4.3.2).

Per introdurre formalmente questo sottoinsieme di m elementi di osservazioni si introduca la seguente notazione:

Sia

- $\Gamma = \{1,2, \ldots\ldots, n\}$
- $\aleph$ è un insieme di m elementi sottoinsieme di Γ
- $\mathbf{1}_m$ è il vettore m-dimensionale di 1
- $H = \{h \in \aleph\ |\text{rank}\ X(\bar{h}) = \text{m}\ \}$

Per ogni $h \in \aleph$ esiste un elemento complementare $\bar{h} = \Gamma - h$; entrambi servono a definire una partizione di y e di X. Quindi, per esempio $y(h)$, definisce un vettore di m elementi $\{y_k : k \in h\}$, mentre $X(\bar{h})$ denota una matrice $(n - m)\ x\ m$ con righe $\{x_k : k \in \bar{h}\}$.

Se la design matrix ha rango m allora l'insieme soluzione $B^*(\theta)$, avrà almeno un elemento della forma:

$$b(h) = X(h)^{-1}y(h) \qquad (4.3.3)$$

essa rappresenta le soluzioni di base che passino per i punti $\{(x_k, y_k), k \epsilon h\}$.

4.4. LA CONDIZIONE DI UNICITÀ DELLA SOLUZIONE

Esistono ovviamente molte soluzioni della forma (4.3.3), per l'esattezza $\binom{n}{m}$ (le $\binom{5}{2}$ =10 rette del caso di Boscovich in Figura 8) quello che il metodo del simplesso fa è cercare la soluzione muovendosi di vertice in vertice considerando la direzione di "steepest descent" (discesa del gradiente).

A questo punto è fondamentale introdurre il teorema che garantisce l'unicità della soluzione (4.3.3) del problema di minimizzazione (4.3.2).

Se *F* è continua allora $\beta^*(h) = X(h)^{-1}y(h)$ è una soluzione unica del problema (4.3.2), se e solo se:

$$(\theta - 1)\mathbf{1}'_{\boldsymbol{m}} < \sum_{k\epsilon\bar{h}} \left[\frac{1}{2} - \frac{1}{2}sgn(y_k - \boldsymbol{x}'_k\beta^*_\theta) - \theta\right]\boldsymbol{x}'_k X(h)^{-1} < \theta\mathbf{1}'_{\boldsymbol{m}} \qquad (4.4.1)$$

Per dimostrarlo introduciamo la derivata direzionale della funzione $\psi(a)$ nella direzione ω:

$$\nabla\psi(\beta_\theta, \omega) = -\sum_{k=1}^{n} \varphi^*(y_k - \boldsymbol{x}'_k\beta_\theta, -\boldsymbol{x}'_k\omega)\boldsymbol{x}'_k\omega, = \sum_{k=1}^{n}\left[\frac{1}{2} - \frac{1}{2}sgn^*(y_k - \boldsymbol{x}'_k\beta_\theta; -\boldsymbol{x}'_k\omega) - \theta\right]\boldsymbol{x}'_k\omega$$

Dove

$$sgn^*(u; z) = \begin{cases} sgn\, u & u \neq 0 \\ sgn\, z & u = 0 \end{cases}$$

Per mostrarlo, se $y_k > x_k\beta_\theta$,

$$-\varphi^*(y_k - \boldsymbol{x}'_k\beta_\theta, -\boldsymbol{x}'_k\omega) = -\theta,$$

$$\frac{1}{2} - \frac{1}{2}sgn^*(y_k - \boldsymbol{x}'_k\beta_\theta; -\boldsymbol{x}'_k\omega) - \theta = -\theta,$$

se $y_k < x_k\beta_\theta$,

$$-\varphi^*(y_k - \boldsymbol{x}'_k\beta_\theta, -\boldsymbol{x}'_k\omega) = 1 - \theta,$$

$$\frac{1}{2} - \frac{1}{2}sgn^*(y_k - \boldsymbol{x}'_k\beta_\theta; -\boldsymbol{x}'_k\omega) - \theta = 1 - \theta,$$

se $y_k = x_k\beta_\theta$,

$$-\varphi^*(y_k - \boldsymbol{x}'_k\beta_\theta, -\boldsymbol{x}'_k\omega) = \begin{cases} 1 - \theta & -\boldsymbol{x}'_k\omega < 0 \\ -\theta & -\boldsymbol{x}'_k\omega \geq 0 \end{cases}$$

$$\left[\frac{1}{2} - \frac{1}{2}sgn^*(y_k - \boldsymbol{x}'_k\beta_\theta; -\boldsymbol{x}'_k\omega) - \theta\right] = \begin{cases} 1 - \theta & -\boldsymbol{x}'_k\omega < 0 \\ -\theta & -\boldsymbol{x}'_k\omega \geq 0 \end{cases}$$

Siccome $\psi(\beta_\theta)$ è una funzione convessa, essa ha un minimo in β^* se e solo se $\psi'(\beta^*_\theta, \omega) > 0$, per ogni

$\omega \neq 0$.

Nel punto $\beta^*(h) = X(h)^{-1}y(h)$:

$$\nabla\psi(\beta^*(h),\omega) = \sum_{k\epsilon h}\left[\frac{1}{2}+\frac{1}{2}sgn(\boldsymbol{x}_k'\omega)-\theta\right]\boldsymbol{x}_k'\omega + \sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^*(y_k-\boldsymbol{x}_k'\beta_\theta^*;-x_k\omega)-\theta\right]\boldsymbol{x}_k'\omega.$$

Se consideriamo infatti $k\,\epsilon\,h$ è chiaro che

$$y_k - \boldsymbol{x}_k'\beta_\theta^* = 0,$$

quindi

$$sgn^*(y_k-\boldsymbol{x}_k'\beta_\theta^*;-\boldsymbol{x}_k'\omega) = sgn(-\boldsymbol{x}_k'\omega)$$

Si ponga $v = X(h)\,\omega$,

quindi

$$\boldsymbol{x}_k'\omega = \boldsymbol{x}_k'X(h)^{-1}v.$$

Inoltre siccome per $\omega \neq 0$ possiamo scrivere

$$v_k = \frac{|v_k|}{sgn(v_k)}$$

Si avrà che $\nabla\psi(\beta^*(h),\omega) > 0$ per ogni $w \neq 0$, se e solo se

$$\sum_{k=1}^{m}\left[\left(\frac{1}{2}-\theta\right)v_k+\frac{1}{2}|v_k|\right]+\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^*(y_k-\boldsymbol{x}_k'\beta_\theta^*;-\boldsymbol{x}_k'X(h)^{-1}v)-\theta\right]\boldsymbol{x}_k'X(h)^{-1}v>0 \qquad (4.4.2)$$

Infine si nota che lo spazio generato dalle direzioni $v\,\epsilon\,R^m$, può essere scritto come combinazione lineare dei vettori della base canonica:

$$e_1=\begin{pmatrix}1\\0\\0\\\vdots\\0\end{pmatrix}\quad e_2=\begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix}\ldots\ldots\ldots\ldots\ldots\quad e_m=\begin{pmatrix}0\\0\\0\\\vdots\\1\end{pmatrix}$$

$$-e_1=\begin{pmatrix}-1\\0\\0\\\vdots\\0\end{pmatrix}\quad -e_2=\begin{pmatrix}0\\-1\\0\\\vdots\\0\end{pmatrix}\ldots\ldots\ldots\ldots\quad -e_m=\begin{pmatrix}0\\0\\0\\\vdots\\-1\end{pmatrix}$$

ovvero lo "span" dei vettori della base canonica definisce l'intero spazio vettoriale. Si può affermare, quindi, che la condizione (4.4.2) è valida per ogni $v\,\epsilon\,R^m$ se e solo se è valida per le $2m$ direzioni canoniche

$$\{\pm e_i : i = 1,\ldots\ldots m\}.$$

Dunque per $v = e_i$ si ha:

$$0 < \sum_{k=1}^{m}\left[\left(\frac{1}{2}-\theta\right)+\frac{1}{2}\right]+\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^*(y_k-\boldsymbol{x}_k'\beta_\theta^*;-\boldsymbol{x}_k'X(h)^{-1}v)-\theta\right]\boldsymbol{x}_k'X(h)^{-1} \leftrightarrow$$

$$(\theta-1)\mathbf{1}_{\boldsymbol{m}}' < \sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^*(y_k-\boldsymbol{x}_k'\beta_\theta^*;-\boldsymbol{x}_k'X(h)^{-1}v)-\theta\right]\boldsymbol{x}_k'X(h)^{-1} \qquad (4.4.3)$$

per $v = -e_i$ si ha:

$$0 < \sum_{k=1}^{m}\left[\left(\theta-\frac{1}{2}\right)+\frac{1}{2}\right]-\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^{*}(y_k-\boldsymbol{x}_k'\beta_\theta^{*};-\boldsymbol{x}_k'X(h)^{-1}v)-\theta\right]\boldsymbol{x}_k'X(h)^{-1}\leftrightarrow$$

$$-\theta\mathbf{1}_{\boldsymbol{m}}'<-\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^{*}(y_k-\boldsymbol{x}_k'\beta_\theta^{*};-\boldsymbol{x}_k'X(h)^{-1}v)-\theta\right]\boldsymbol{x}_k'X(h)^{-1}\leftrightarrow$$

$$\theta\mathbf{1}_{\boldsymbol{m}}'>\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^{*}(y_k-\boldsymbol{x}_k'\beta_\theta^{*};-\boldsymbol{x}_k'X(h)^{-1}v)-\theta\right]\boldsymbol{x}_k'X(h)^{-1} \qquad (4.4.4)$$

Combinando la (4.4.3) e la (4.4.4), possiamo definire la condizione:

$$(\theta-1)\mathbf{1}_{\boldsymbol{m}}'<\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn^{*}(y_k-\boldsymbol{x}_k'\beta_\theta^{*};-\boldsymbol{x}_k'X(h)^{-1}v)-\theta\right]\boldsymbol{x}_k'X(h)^{-1}<\theta\mathbf{1}_{\boldsymbol{m}}' \qquad (4.4.5)$$

Se consideriamo il caso di variabile assolutamente continua, per ogni osservazione $k\,\epsilon\,\bar{h}$ succede che:

$P(Y_k=\boldsymbol{x}_k'\beta_\theta^{*})=0$

Quindi scompare la dipendenza da v poichè :

$sgn^{*}(y_k-\boldsymbol{x}_k'\beta_\theta^{*};-\boldsymbol{x}_k'X(h)^{-1}v)=sgn(y_k-\boldsymbol{x}_k'\beta_\theta^{*})$,

e risulta dimostrata la (4.4.1).

Si dirà che le osservazioni (y,X) sono in posizione generale se un numero m di queste, è tale che per ogni $h\in H$:

$y_k-x_kb(h)\neq 0$ per ogni $k\notin h$,

ovvero non deve esserci un numero di residui uguale a zero maggiore di m.

Se (y,X) sono in posizione generale, esiste un'unica soluzione al problema della Quantile Regression della forma: $b(h)=X(h)^{-1}y(h)$, se e solo se è verificata la (4.4.1), ovvero se:

$$(\theta-1)\mathbf{1}_{\boldsymbol{m}}'<\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn(y_k-\boldsymbol{x}_k'\beta_\theta^{*})-\theta\right]\boldsymbol{x}_k'X(h)^{-1}<\theta\mathbf{1}_{\boldsymbol{m}}'$$

Se le disuguaglianze non sono strette allora esisteranno più soluzioni della forma $b(h)$ e l'insieme soluzione corrisponderà all'inviluppo convesso delle molteplici soluzioni della forma $b(h)$.

È interessante considerare cosa succede nel location model (modello con sola intercetta (4.2.4)) in seguito al teorema appena dimostrato.

Se $x_k=1$ con $k=1,\ldots\ldots,n$, allora $H=\Gamma$ e se F è continua, allora il teorema precedente afferma che:

$\beta_\theta^{*}=y(h)$ è l'unico quantile di livello θ se e solo se:

$$(\theta-1)<\sum_{k\epsilon\bar{h}}\left[\frac{1}{2}-\frac{1}{2}sgn\big(y_k-y(h)\big)-\theta\right]<\theta \qquad (4.4.6)$$

L'espressione in parentesi è pari a $-\theta$ se $y_k > y(h)$ e $1-\theta$ se $y_k < y(h)$, quindi se indichiamo con π il numero di volte in cui $y_k < y(h)$ e con $n-\pi-1$ il numero di volte in cui $y_k > y(h)$ allora riscriviamo la condizione (4.4.6):

$$\begin{cases} \theta - 1 < \pi(1-\theta) - (n-\pi-1)\theta & se\ \ y_k > y(h) \\ \theta > \pi(1-\theta) - (n-\pi-1)\theta & se\ \ y_k < y(h) \end{cases} \leftrightarrow$$

$$\begin{cases} \theta - 1 < \pi - \pi\theta - n\theta + \pi\theta + \theta & se\ \ y_k > y(h) \\ \theta > \pi - \pi\theta - n\theta + \pi\theta + \theta & se\ \ y_k < y(h) \end{cases} \leftrightarrow$$

$$\begin{cases} \pi > n\theta - 1 & se\ \ y_k > y(h) \\ \pi < n\theta & se\ \ y_k < y(h) \end{cases}$$

Quindi la condizione (4.4.6) si riduce a richiedere che il numero di osservazioni inferiori a $y(h)$ siano comprese tra $n\theta - 1$ ed $n\theta$.

Come visto in precedenza, la non unicità della soluzione, può quindi presentarsi per due motivazioni:

- Presenza di osservazioni ripetute e quindi un numero di residui maggiori di zero superiori a m. Tale caso è inusuale, ma possibile solo se la Y è discreta.
- Oppure nel caso in cui $n\theta$ sia una quantità intera, in quanto π, essendo per forza intero, non potrà mai essere strettamente compreso tra le due quantità intere successive $n\theta - 1$ e $n\theta$.

Si conclude quindi che nella Quantile Regression, se la variabile risposta è assolutamente continua, allora la soluzione $\beta^*(h) = X(h)^{-1}y(h)$ è unica se e solo se è verificata la (4.4.1). Considerando invece il location model, (caso particolare di QR) la condizione per l'unicità della soluzione si può sintetizzare richiedendo soltanto che $n\theta$ sia non intero.

4.5. INFERENZA E RISULTATI ASINTOTICI

4.5.1. LA DISTRIBUZIONE DI PROBABILITÀ DELLO STIMATORE DEI PARAMETRI

Per introdurre le procedure di inferenza sulla Quantile Regression, si inizierà definendo un apparato sul location model (modello con sola intercetta per la definizione del quantile non condizionato) e su campioni finiti, per poi passare ai quantili condizionati e alla teoria asintotica.

Si supponga che $Y_1, Y_2, \dots\dots\dots\dots, Y_n$ siano variabili aleatorie indipendenti e identicamente distribuite con funzione di distribuzione F, e si assuma che F abbia una densità f continua in un intorno del quantile di livello θ:

$Q_\theta = F^{-1}(\theta)$, con $f(Q_\theta) > 0$.

Considerando il modello (4.2.3), il quantile stimato sarà:

$\hat{Q}_\theta = inf_a\{a \in R | \sum \rho_\theta(Y_k - a) = min\}$.

Si nota che la funzione obiettivo è convessa in quanto somma di funzioni convesse, per cui il gradiente della funzione obiettivo:

$g_n(a) = \sum_{k=1}^{n}(I(Y_k < a) - \theta)$,

è monotono.

Dalla monotonicità del gradiente accade che:

$$P\left(\hat{Q}_\theta > a\right) = P(g_n(a) < 0) = P(\sum_{k=1}^{n} I(Y_k < a) < n\theta) = P\left(B\left(n, F(a)\right) < n\theta\right) \quad (4.5.1)$$

Dove $B(n,p)$ denota una variabile binomiale di parametri (n,p)

Per comprendere meglio il senso della (4.5.1) si supponga di avere a disposizione un campione di 101 elementi della variabile Y.

Il quantile campionario per $\theta = 0{,}5$ corrisponderà alla 51-esima posizione del campione ordinato, quindi esso risulterà maggiore di un generico $a \in R$, se e solo se il numero di volte in cui Y è inferiore ad a sarà inferiore a $n\theta = 101 * 0{,}5 = 50{,}5$, per cui è chiaro che:

$$P\left(\hat{Q}_\theta > a\right) = P(g_n(a) < 0)$$

Infatti il verificarsi di questo evento darà la certezza che la 51-esima posizione risulterà comunque superiore ad a, se ad esempio $g_n(a) = \sum_{k=1}^{n} I(y_k < a) - 50{,}5 = q - 50{,}5 < 0$, significa che la $(q+1)$-esima posizione del campione ordinato avrà un valore più alto di a. Poiché q è nel migliore dei casi 50, allora la 51-esima osservazione sarà con certezza maggiore di a.

Indicando con:

$s = \lceil n\theta \rceil$,

il più piccolo intero maggiore o uguale di $n\theta$, si può esprimere la funzione di distribuzione dello stimatore $\hat{Q}_\theta$: $G_{\hat{Q}_\theta}(a) = P\left(\hat{Q}_\theta \leq a\right)$, usando la funzione beta incompleta:

$$G_{\hat{Q}_\theta}(a) = 1 - \sum_{k=s}^{n}\binom{n}{k}F(a)^k\left(1 - F(a)\right)^{n-k} = n\binom{n-1}{s-1}\int_0^{F(a)} t^{s-1}(1-t)^{n-s}dt\ .$$

Derivando ambo i membri si ottiene la funzione densità di $\hat{Q}_\theta$:

$$g_{\hat{Q}_\theta}(a) = n\binom{n-1}{s-1}F(a)^{s-1}\left(1 - F(a)\right)^{n-s}f(a). \quad (4.5.2)$$

Per comprendere meglio la struttura della (4.5.2), si nota che una densità di questo tipo, indica che l'evento $\left\{x < Y_{(s)} < x + \delta\right\}$, richiede che $s-1$ osservazioni siano minori di x e $n-s$ maggiori di $x+\delta$, il numero di combinazioni con cui tale risultato può essere ottenuto, è proprio: $n\binom{n-1}{s-1}$ e ogni combinazione ha una probabilità pari a $F(a)^{s-1}\left(1 - F(a)\right)^{n-s}[F(a+\delta) - F(a)]$.

Infatti è possibile dimostrare la (4.5.2) partendo dalla definizione:

$$P\{a < Y_{(s)} < a+\delta\} = n\binom{n-1}{s-1}F(a)^{s-1}\big(1-F(a)\big)^{n-s}[F(a+\delta)-F(a)] \leftrightarrow$$

$$P\{a < Y_{(s)} < a+\delta\} = n\binom{n-1}{s-1}F(a)^{s-1}\big(1-F(a)\big)^{n-s}[f(a)\delta + o(\delta^2)]$$

Dividendo ambo i membri per δ e considerando il limite per δ che tende a zero si ottiene la (4.5.2).

Questo approccio può anche essere utilizzato per costruire intervalli di confidenza per Q_θ nella forma,

$$P\{\hat{Q}_{\theta_1} < Q_\theta < \hat{Q}_{\theta_2}\} = 1-\alpha, \tag{4.5.3}$$

dove θ_1 e θ_2 sono scelti in modo da soddisfare la relazione:

$$P\{n\theta_1 < B(n,\theta) < n\theta_2\} = 1-\alpha \tag{4.5.4}$$

Tale metodo di stima di intervalli di confidenza è detto **metodo diretto.**

Una volta stabilita la distribuzione di probabilità del quantile campionario, considerato come soluzione di minimo del location model, in campioni finiti, occorre definire la densità dello stimatore Quantile Regression β_θ^*.

Si consideri il modello lineare:

$$Y_k = x_k'\beta + u_k, \quad k = 1, \ldots\ldots, n$$

con errori u_k i.i.d., aventi la stessa funzione di distribuzione F (essa rappresenta la funzione di distribuzione degli errori) e densità strettamente positiva f nel punto $F^{-1}(\theta)$. Allora la densità di β_θ^* assumerà la forma:

$$g(b) = \sum_{h\in\aleph} P\{D_h(b) \in C\} \cdot |X(h)| \cdot \prod_{k\in h} f\big(\boldsymbol{x}_k'(b-\beta_\theta) + F^{-1}(\theta)\big), \tag{4.5.5}$$

dove:

- $D_h(b) = \sum_{k\in\bar{h}}\left[\frac{1}{2} - \frac{1}{2}sgn(y_k - \boldsymbol{x}_k'b) - \theta\right]\boldsymbol{x}_k' X(h)^{-1}$
- C denota l'ipercubo $[\theta-1, \theta]^m$

Sappiamo dalla condizione (4.4.1), che $\beta_\theta^* = b(h) = X(h)^{-1}y(h)$ se e solo se $D_h(b(h)) \in C$.

Per $b \in R^m$, $B(b,\delta) = b + [-\delta/2, \delta/2]^m$ definisce l'ipercubo centrato in b con lati di lunghezza δ e sia:

$$P\{\beta_\theta^* \in B(b,\delta)\} = \sum_{h\in\aleph} P\big\{b(h) \in B(b,\delta), D_h\big(b(h)\big) \in C\big\} \leftrightarrow$$

$$P\{\beta_\theta^* \in B(b,\delta)\} = \sum_{h\in\aleph} E\{I[b(h) \in B(b,\delta)]\}P\{D_h(b(h)) \in C | b(h) \in B(b,\delta))\}$$

Siccome $h \in \aleph$ allora condizionare all'evento $b(h) \in B(b,\delta)$ (il che implica conoscere l'intervallo in cui cade $b(h)$) equivale a condizionare rispetto a $Y(h)$ in quanto se $h \in \aleph$

$$X(h)b(h) = X(h)X(h)^{-1}y(h) = y(h).$$

Si può quindi concludere che:

$$P\{\beta_\theta^* \in B(b,\delta)\} = \sum_{h\in\aleph} E\{I[b(h) \in B(b,\delta)]\}P\{D_h(b(h)) \in C|Y(h))\}. \quad (4.5.6)$$

La distribuzione di probabilità di $D_h\big(b(h)\big)$ è discreta e presenta un numero di determinazioni pari a 2^{n-m}.

Per comprenderlo basta pensare al caso uni-dimensionale, $(m = 1)$ in cui $D_h\big(b(h)\big)$ assume la forma:

$$D_h\big(b(h)\big) = \sum_{k\epsilon\bar{h}} \left[\frac{1}{2} - \frac{1}{2}sgn\big(Y_k - y(h)\big) - \theta\right] x_k X(h)^{-1}$$

In questo caso, $\bar{h}$ ha $n-1$ elementi, quindi si avranno $n-1$ valori di y_k, ciascuno dei quali potrà essere superiore o inferiore alla singola $y(h)$ e fornire 2 diversi valori alla quantità $sgn\big(Y_k - y(h)\big)$. È come se si dovessero contare il numero di possibili esiti di una partita a testa o croce su n-1 partite, che sono in tutto 2^{n-1}.

Dividendo ambo i membri per δ^m ed eseguendo il limite per $\delta \to 0$, il termine di sinistra della (4.5.6) rappresenterà la densità di β_θ^*, mentre nel membro di destra la probabilità $P\{D_h(b(h)) \in C|Y(h))\} \to P\{D_h(b) \in C)\}$ e il termine $\frac{E\{I[b(h)\in B(b,\delta)]\}}{\delta^m} = \frac{P\{b(h)\in B(b,\delta)\}}{\delta^m}$ tenderà alla densità di $b(h)$, ovvero alla densità congiunta di $X(h)^{-1}Y(h)$.

Poiché si può scrivere:

$$f_{Y(h)}(y) = \prod_{k\in h} f(y_k - \boldsymbol{x}_k'\beta_\theta)$$

Allora la densità congiunta di $X(h)^{-1}Y(h)$ può scriversi come:

$$f_{X(h)^{-1}Y(h)}(b) = |X(h)|\prod_{k\in h} f\big((X(h)b)_k - \boldsymbol{x}_k'\beta_\theta + F^{-1}(\theta)\big) = |X(h)|\prod_{k\in h} f\big(\boldsymbol{x}_k'(b-\beta_\theta) + F^{-1}(\theta)\big)$$

Sostituendo gli elementi nell'equazione (4.5.6) si ottiene la ((4.5.5)).

Si noti che i valori di h per cui la matrice $X(h)$ è singolare non contribuiscono alla densità in quanto $|X(h)| = 0$.

A questo punto, una volta definite le distribuzioni degli stimatori nel caso di campioni finiti, si definisce la loro approssimazione asintotica. Come in precedenza, si ragionerà inizialmente sulla distribuzione asintotica del quantile campionario, soluzione del problema di minimo (4.2.3)

Ricordando che la funzione obiettivo (4.2.3) è convessa, allora il suo gradiente sarà tale che:

$$\bar{g}_n(a) = n^{-1}g_n(a) = n^{-1}\sum_{k=1}^n (I(Y_k < a) - \theta),$$

Dalla monotonicità del gradiente è chiaro che $\hat{Q}_\theta > a$ se e solo se $\bar{g}_n(a) < 0$, si avrà dunque:

$$P\big(\sqrt{n}\big(\hat{Q}_\theta - Q_\theta\big) > \delta\big) = P\left(\hat{Q}_\theta - Q_\theta - \frac{\delta}{\sqrt{n}} > 0\right) = P\left(\bar{g}_n\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right) < 0\right) =$$

$$P\left(n^{-1}\sum_{k=1}^{n}\left(I\left(Y_k < Q_\theta + \frac{\delta}{\sqrt{n}}\right) - \theta\right) < 0\right)$$

Si è ridotto quindi il comportamento di $\hat{Q}_\theta$ ad un problema di limite centrale DeMoivre-Laplace, in cui si ha un array triangolare di variabili bernoulliane.

Gli addendi della somma assumono valori $(1-\theta)$ e $-\theta$ con probabilità $F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)$ e $1 - F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)$; dato che per n sufficientemente grande:

$$E\left[\bar{g}_n\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)\right] = E\left[n^{-1}\sum_{k=1}^{n}\left(I\left(Y_k < Q_\theta + \frac{\delta}{\sqrt{n}}\right) - \theta\right)\right] = n^{-1}\left[nF\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right) - n\theta\right] = F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right) - \theta$$

$$= F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right) - \theta = F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right) - F(Q_\theta) = f(Q_\theta)\frac{\delta}{\sqrt{n}} + o\left[\left(\frac{\delta}{\sqrt{n}}\right)^2\right] \cong f(Q_\theta)\frac{\delta}{\sqrt{n}}$$

e

$$V\left[\bar{g}_n\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)\right] = V\left[n^{-1}\sum_{k=1}^{n}\left(I\left(Y_k < Q_\theta + \frac{\delta}{\sqrt{n}}\right) - \theta\right)\right] = n^{-2}\left[nF\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)\left(1 - F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)\right)\right]$$

$$= n^{-1}\left[F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)\left(1 - F\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right)\right)\right] \cong \frac{\theta(1-\theta)}{n}$$

Allora possiamo porre $\omega^2 = \frac{\theta(1-\theta)}{f^2(Q_\theta)}$ e scrivere:

$$P\left(\sqrt{n}\left(\hat{Q}_\theta - Q_\theta\right) > \delta\right) = P\left(\frac{g_n\left(Q_\theta + \frac{\delta}{\sqrt{n}}\right) - f(Q_\theta)\frac{\delta}{\sqrt{n}}}{\sqrt{\frac{\theta(1-\theta)}{n}}} < -\omega^{-1}\delta\right) \cong 1 - \phi(\omega^{-1}\delta)$$

e dunque

$$\sqrt{n}\left(\hat{Q}_\theta - Q_\theta\right) \sim N(0, \omega^2).$$

Dall'ultima relazione, non è soltanto evidente che la distribuzione asintotica dello stimatore è normale, ma è anche chiara la non distorsione asintotica.

Estendendo questo risultato al caso di distribuzione congiunta di più quantili, si ponga: $\hat{\zeta}_n = \left(\hat{Q}_{\theta_1}, \ldots., \hat{Q}_{\theta_s}\right)$, con $\zeta_n = \left(Q_{\theta_1}, \ldots., Q_{\theta_s}\right)$ e si ottiene:

$$\sqrt{n}\left(\hat{\zeta}_n - \zeta_n\right) \sim N(0, \Omega)$$

dove $\Omega = \omega_{ij} = \left(\theta_i \wedge \theta_j - \theta_i\theta_j\right) / \left(f\left(F^{-1}(\theta_i)\right) f\left(F^{-1}(\theta_j)\right)\right)$ (4.5.7)

Questi risultati sul caso del quantile ordinario, possono essere generalizzati al modello di regressione lineare

$$Y_k = \boldsymbol{x}_k'\beta + u_i.$$

Si ipotizza, inoltre, che $\sum x_k \boldsymbol{x}_k' \equiv W_n$ converge, al crescere di n, a una matrice definita positiva W_0.

Allora la distribuzione asintotica congiunta degli s m-variati stimatori della Quantile Regression $\hat{\zeta}_n = \left(\boldsymbol{\beta}^*_{\theta_1}{}', \ldots\ldots, \boldsymbol{\beta}^*_{\theta_s}{}'\right)'$, assume la forma:

$$\sqrt{n}\left(\hat{\zeta}_n - \zeta_n\right) = \left(\sqrt{n}\left(\beta^*_{\theta_i} - \beta^*_{\theta_j}\right)\right)_{j=1}^{m} \sim N(0, \Omega \otimes W_0^{-1}).$$

Osservando la (4.5.7), risulta evidente che la precisione delle stime della Quantile Regression, dipenda dal reciproco della funzione di densità valutata nel quantile di interesse $\frac{1}{f\left(F^{-1}(\theta)\right)} = s(\theta)$. Tale funzione prende il nome di "sparsity function", ed è ovvio che sia così, in quanto essa rappresenta la densità delle osservazioni intorno al quantile di interesse. Quando la sparsity function è piccola, le osservazioni si addensano molto intorno al quantile, e la stima risulterà migliore.

Per conoscere la qualità della nostra stima, dunque, occorrerà entrare nel campo dello 'smoothing' di funzioni e della stima delle funzioni di densità. In teoria, ciò potrebbe essere evitato introducendo tecniche simulative come il bootstrapping, ma è interessante osservare come esistano tecniche dirette per il calcolo della matrice asintotica delle covarianze. Differenziando l'identità $F\left(F^{-1}(t)\right) = t$ è chiaro che la sparsity function è la derivata della funzione quantile; infatti:

$$\frac{d}{dt}F^{-1}(t)f\left(F^{-1}(t)\right) = 1 \leftrightarrow \frac{d}{dt}F^{-1}(t) = s(t)$$

Quindi differenziando, se possibile, la funzione di distribuzione F, si ottiene la funzione di densità f e differenziando la funzione quantile F^{-1}, si ottiene la sparsity function s. Risulta quindi immediato, stimare la sparsity function usando il rapporto incrementale della funzione quantile empirica:

$$\hat{s}_n(t) = \frac{\left[\hat{F}_n^{-1}(t+h_n) - \hat{F}_n^{-1}(t-h_n)\right]}{2h_n}$$

dove $\hat{F}_n^{-1}$è una stima di F^{-1} ed h_n è l'ampiezza di un intervallo (bandwith) che tende a zero per $\to \infty$. Una regola di bandwith suggerita da Hall e Sheather (1988) è basata sull'espansione di Edgeworth per quantili Studentized ed è:

$$h_n = n^{-1/3} z_\alpha^{2/3} [1{,}5 s(t)/s''(t)]^{1/3}$$

Dove z_α è il quantile di livello $(1 - \alpha/2)$ di una normale standard.

In assenza di altre informazioni sulla forma di $s(\cdot)$, si può usare un modello Gaussiano per definire la forma di h_n, che determina:

$$h_n = n^{-1/3} z_\alpha^{2/3} [1{,}5\varphi^2(\Phi^{-1}(t))/(2(\Phi^{-1}(t))^2 + 1)]^{1/3}$$

Avendo scelto il bandwith h_n il prossimo punto è come scegliere $\hat{F}^{-1}$; l'approccio più semplice dovrebbe essere quello di utilizzare i residui del fitting della Quantile Regression. Siano:

$r_k : k = 1, \ldots\ldots, n$ il vettore dei residui

$r_{(k)}: k = 1, \dots \dots, n$ le corrispondenti statistiche ordinate,

allora è possibile definire la funzione quantile empirica $\hat{F}^{-1}(t) = r_{(j)}$, con $t \in \left[\frac{(j-1)}{n}, \frac{j}{n}\right]$.

Un possibile svantaggio nella stima della sparsity function tramite i residui è che se si stima un modello in cui il numero dei parametri m è prossimo a n, allora si avranno un numero di residui nulli almeno pari a m e il bandwith dovrà essere abbastanza grande da garantire che questi zeri vengano evitati. Si fa notare che, nel caso di variabili discrete, tale problema si acuisce ancor di più in quanto i residui nulli potrebbero essere più di m.

L'approccio più semplice, nella stima di $\hat{F}^{-1}$ è quello di non considerare questi residui nulli ed effettuare i calcoli direttamente su un campione di $n - m$ residui.

Il problema della stima della sparsity function non è un problema di facile soluzione e diventa cruciale, non solo per il calcolo delle distribuzioni asintotiche, ma anche in alcuni tipi di intervalli di confidenza.

4.5.2. INTERVALLI DI CONFIDENZA

Il fatto che si siano definite le distribuzioni di probabilità degli stimatori dei parametri β_θ, ci fa dedurre che esistano tecniche di calcolo di intervalli di confidenza per questi stessi parametri. Si farà notare, nell' applicazione, che gli intervalli di confidenza di β_θ, sono nella maggior parte dei casi, meno volatili rispetto a quelli derivanti dai GLM.

Si è già considerata nei paragrafi precedenti la tecnica di stima degli intervalli di confidenza denominata "metodo diretto", definita dalle relazioni (4.5.3) e (4.5.4); tuttavia esistono ulteriori vie che, invece, verranno descritte in questo paragrafo, sottolineandone vantaggi e svantaggi relativi.

Tra le tecniche di calcolo per intervalli di confidenza dei quantili, particolare attenzione la merita lo studentization approach (Efron 82, Hall 88). Per comprendere la ratio sottostante tale approccio, si farà riferimento al location model e si ipotizzerà un campione casuale $Y_1, \dots \dots, Y_n$ di una popolazione con funzione di distribuzione F, densità f e quantile F^{-1}. Sia $Y_{(1)}, \dots \dots, Y_{(n)}$ il corrispondente campione ordinato e siano $\hat{F}_n$ ed $\hat{F}_n{}^{-1}$ rispettivamente le funzioni di ripartizione e quantile empiriche. L'intervallo di confidenza per $F^{-1}(\theta)$, con $\theta \in (0,1)$, ha la forma standard:

$$\hat{F}_n{}^{-1}(\theta) \pm s z_\alpha \sqrt{\theta(1-\theta)/n}\,,$$

dove s è uno stimatore consistente (a livello $n^{\frac{1}{4}}$) della quantità di $1/f\left(F^{-1}(\theta)\right)$, ovvero della sparsity function.

Per ottenere quindi un intervallo di confidenza per $F^{-1}(\theta)$ bisogna ottenere uno stimatore consistente della sparsity function, il che, come già detto, non risulta semplice.

Il metodo bootstrap permetterebbe di aggirare il problema della stima di $1/f\left(F^{-1}(\theta)\right)$, ma è altresì vero che molti autori considerano tale tecnica non soddisfacente nella definizione di intervalli di confidenza per i quantili.

Sia l'approccio studentized, che quello diretto possono essere generalizzati al caso della Quantile Regression; si definiscono le generalizzazioni studentized e diretta, degli intervalli di confidenza a livello $(1-2\alpha)$, al caso Quantile Regression:

- Studentized: $I_{1n} = (\boldsymbol{X}\beta_\theta - a_n, \boldsymbol{X}\beta_\theta + a_n)$
  Dove :
  - $a_n = sz_\alpha\sqrt{\boldsymbol{X'Q^{-1}X}\theta(1-\theta)/n}$
  - s è uno stimatore consistente di $1/f\left(F^{-1}(\theta)\right)$
  - z_α è il quantile di livello $1-\alpha$ di una normale standard

- Diretto: $I_{2n} = \left(\boldsymbol{X}\beta_{\theta-b_n}, \boldsymbol{X}\beta_{\theta-b_n}\right)$
  Dove :
  - $b_n = z_\alpha\sqrt{\boldsymbol{X'Q^{-1}X}\theta(1-\theta)/n}$
  - z_α è il quantile di livello $1-\alpha$ di una normale standard

Date le seguenti condizioni:

- F è due volte differenziabile nel punto $F^{-1}(\theta)$ ed $f\left(\hat{F}^{-1}(\theta)\right) = F'\left(\hat{F}^{-1}(\theta)\right)$
- $max_{i,j}|x_{i,j}| = O\left(n^{1/4}\right)$
- $\sum_{i=1}^{n}\left\|x_{i,j}\right\|^3 = O(n)$
- $n^{-1}\sum_{i=1}^{n} x_i\, x_i' = Q + O\left(n^{-1/4}\log n\right)$ dove Q è una matrice $p\ x\ p$ definita positiva

è possibile definire le proprietà asintotiche degli stimatori intervallari I_{1n} ed I_{2n} , infatti per $n \to \infty$ risulta:

- $\boldsymbol{X}\hat{\beta}_\theta \pm a_n = \boldsymbol{X}\hat{\beta}_{\theta-b_n} + O_p\left(\left(\frac{\log n}{n}\right)^{3/4}\right)$
- $\sqrt{\boldsymbol{n}}\left(\boldsymbol{X}\hat{\beta}_{\theta-b_n} - \boldsymbol{X}\beta_\theta\right)\vec{\boldsymbol{d}}\ \boldsymbol{N}\left(\pm\frac{z_\alpha\sqrt{\boldsymbol{X'Q^{-1}X}\theta(1-\theta)/n}}{f\left(F^{-1}(\theta)\right)}, \frac{\boldsymbol{X'Q^{-1}X}\theta(1-\theta)}{f^2\left(F^{-1}(\theta)\right)}\right)$
- $P[\boldsymbol{X}\beta_\theta \in I_{1n}] = P[\boldsymbol{X}\beta_\theta \in I_{2n}] + O\left(n^{-1/4}(\log n)^{3/4}\right) = 1 - 2\alpha + O\left(n^{-1/4}(\log n)^{3/4}\right)$

Per concludere, si introduce l'ultima tecnica di stima degli intervalli di confidenza, che si ispira alla teoria della stima intervallare non parametrica, in particolare alla teoria dei rank tests.

La teoria classica dei rank tests come sviluppata da Hajek e Sidak (1967) inizia con l'introduzione della rank score function (4.2.6), che per comodità riportiamo:

$$\hat{a}_{nk}(t) = \begin{cases} 1 & if\ t \le (R_k - 1)/n \\ R_k - tn & if (R_k - 1)/n < t \le R_k/n \\ 0 & if\ t > R_k/n \end{cases}$$

dove R_k è la posizione della k-esima osservazione di Y. Integrando $\hat{a}_{nk}(t)$ rispetto a differenti funzioni score generating φ, possiamo ottenere alcune statistiche di ranking, che possono essere utilizzate per effettuare test.

Ad esempio integrando rispetto alla misura di Lebesgue otteniamo gli Wilcoxon scores:

$$b_k = \int_0^1 \hat{a}_{nk}(t)dt = \frac{\left(R_k - \frac{1}{2}\right)}{n} : k = 1, \dots, n$$

A questo punto, data l'invarianza dell'ordinamento per trasformazioni monotone, si può affermare che gli R_k possono essere visti come le posizioni di un campione uniforme $\{U_1, \dots\dots, U_n\}$, con $U_k = F(Y_k)$ e la funzione rankscore $\hat{a}_{nk}(t)$, può essere interpretata come l'equivalente empirico di $I(Y_k > F^{-1}(t)) = I(U_k > t)$.

Considerando $(c_{1n}, \dots\dots, c_{nn})$ come un array triangolare di numeri reali, che soddisfino la condizione di Lundberg:

$$\frac{\max(c_{kn} - \bar{c}_n)^2}{\sum_{k=1}^n (c_{kn} - \bar{c}_n)^2} \to 0.$$

Dove $\bar{c}_n = n^{-1} \sum c_{kn}$ e si assuma che $\{Y_1, \dots\dots, Y_n\}$, costituisce un campione casuale da una variabile aleatoria con funzione di distribuzione F assolutamente continua. Allora per il teorema di Donsker:

$$Z_n(t) = [\sum_{k=1}^n (c_{kn} - \bar{c}_n)^2]^{-1/2} \sum_{k=1}^n (c_{kn} - \bar{c}_n)\hat{a}_{kn}$$

converge a un ponte Browniano su $C[0,1]$.

È stata sviluppata, sulla scorta di questi risultati, una teoria per una classe di statistiche lineari di ranking della forma:

$$S_n = [\sum_{k=1}^n (c_{kn} - \bar{c}_n)^2]^{-1/2} \sum_{k=1}^n (c_{kn} - \bar{c}_n)\hat{b}_k\ ,$$

dove:

$$\hat{b}_k = -\int \varphi(t)\, d\hat{a}_k(t),$$

$$\varphi(t) = \theta - I\{t < \theta\}.$$

In particolare, per funzioni quadrato integrabili φ: $[0,1] \to R$ abbiamo la rappresentazione lineare

$$S_n = [\sum_{k=1}^n (c_{kn} - \bar{c}_n)^2]^{-1/2} \sum_{k=1}^n (c_{kn} - \bar{c}_n)\varphi(U_k) + o_p(1),$$

e conseguentemente S_n è asintoticamente Gaussiana sotto l'ipotesi nulla, con media zero e varianza $A^2(\varphi) = \int(\varphi(t) - \bar{\varphi})^2 dt$,

dove $\bar{\varphi} = \int \varphi(t)dt$.

Ora il punto è come possa quest'idea essere estesa alla regressione. A questa domanda risposero Gutenbunner e Jureckova (1993), che osservarono che le funzioni di rankscores possono essere viste come un caso speciale di una formulazione più generale per il modello lineare, in cui la funzione $\hat{a}_{nk}$ è definita in termini del problema di programmazione lineare:

$max\{y'a|X'a = (1-t)X'1, a \in [0,1]^m\}$

Questo è il problema duale introdotto con la (4.2.5). Come sviluppato da Gutenbunner e Jureckova, il test d'ipotesi $\beta_2 = 0 \in R^q$ nel modello $y = X_1\beta_1 + X_2\beta_2 + u$, basato sul processo di rankscore, può essere costruito:

- calcolando $\{\hat{a}_{nk}(t)\}$ sul modello ristretto: $y = X_1\beta_1 + u$,
- calcolando il vettore n-dimensionale b, con elementi $b_k = -\int \varphi(t)d\hat{a}_{nk}(t)$,
- formando il vettore q-dimensionale $S_n = n^{-1/2}X_2'b$.

È chiaro che sotto l'ipotesi nulla sarà:

$S_n \sim N(0, A^2(\varphi)Q_0)$,

con:

- $A^2(\varphi) = \int_0^1 \varphi^2(t)dt$,
- $Q_0 = \lim_{n\to\infty} Q_n$,
- $Q_n = \left(X_2 - \hat{X}_2\right)'\left(X_2 - \hat{X}_2\right)/n$,
- $\hat{X}_2 = X_1(X_1'X_1)^{-1}X_1'X_2$.

Così, la statistica test $T_n = \frac{S_n'{Q_0}^{-1}S_n}{A^2(\varphi)}$ si distribuisce asintoticamente come una χ_q^2.

Utilizzando la funzione

$\varphi_\theta(t) = \theta - I(t < \theta)$, e procedendo come descritto sopra, si trova che:

$\hat{b}_{nk} = -\int_0^1 \varphi_\theta(t)d\hat{a}_{nk}(t) = \hat{a}_{nk}(\theta) - (1-\theta)$,

$A^2(\varphi_\theta) = \int_0^1(\varphi_\theta(t) - \bar{\varphi})^2 dt = \theta(1-\theta)$.

Quindi un test d'ipotesi con ipotesi nulla $H_0: \beta_2 = Q$, si baserà su $\hat{a}_n$, risolvendo:

$max\{(y - x_2Q)'a|X_1'a = (1-t)X_1'1, a \in [0,1]^n\}$, (4.5.8)

con $S_n = n^{-\frac{1}{2}} X_2' \hat{b}_n(Q) \sim N(0, A^2(\varphi_\theta) q_n^2)$ sotto H_0, dove:

$q_n^2 = n^1 x_2'(I - X_1(X_1'X_1)^{-1}X_1')x_1.$

A questo punto si può calcolare la statistica test:

$T_n(Q) = S_n'(Q)/(A(\varphi_\theta)q_n)$

E rigettare l'ipotesi nulla se:

$|T_n(Q)| > \phi^{-1}(1 - \alpha/2).$

Un intervallo di confidenza definito in tal modo non è simmetrico, ma ha il grande vantaggio di non necessitare della stima della sparsity function.

4.6. ALCUNE PROPRIETÀ

Dopo l'introduzione basilare sulla regressione del quantile si definiscono alcuni risultati, e le proprietà più rilevanti, che si osservano e che la rendono particolarmente interessante anche rispetto al modello di regressione lineare classico. Come già detto, il modello di regressione lineare del quantile, per qualsiasi valore di $\theta \in (0,1)$, permette di stimare i quantili condizionati di Y data una matrice di regressori X, dove al variare di θ è possibile valutare l'intera distribuzione della variabile risposta condizionatamente alle esplicative, risultato nettamente differente da una semplice valutazione della media condizionata.

Valgono inoltre:

$$\beta_\theta^*(\lambda y, x) = \lambda \beta_\theta^*(y, x) \quad \lambda \in (0, \infty) \tag{4.6.1}$$

$$\beta_\theta^*(-\lambda y, x) = \lambda \beta_{1-\theta}^*(y, x) \quad \lambda \in (0, \infty) \tag{4.6.2}$$

$$\beta_\theta^*(y + x\lambda, x) = \beta_\theta^*(y, x) + \lambda \quad \lambda \in (0, \infty) \tag{4.6.3}$$

$$\beta_\theta^*(y, xA) = A^{-1}\beta_\theta^*(y, x) \quad A \text{ è una matrice invertibile} \tag{4.6.4}$$

Dove le equazioni (4.6.1) e (4.6.2) definiscono equivarianze di scala: se la risposta y è riscalata per un fattore λ, allora il vettore dei parametri subisce la medesima operazione di scala. La proprietà (4.6.3) è chiamata proprietà di equivarianza della regressione o di shift, infine la proprietà (4.6.4) è detta equivarianza alle riparametrizzazioni del disegno e significa che combinazioni lineari dei regressori si trasferiscono sul vettore dei parametri stimati con l'inversa della matrice di trasformazione. Alcune di queste proprietà sono presenti anche nello stimatore dei minimi quadrati ordinari, tuttavia nella regressione del quantile vi è una proprietà ben più forte delle proprietà di equivarianza già discusse, che non viene solitamente condivisa dagli altri tipi di regressione.

Presa una qualunque funzione a valori reali strettamente monotona g, segue che:

$$\hat{Q}_\theta(g(Y)|\boldsymbol{x}_k) = g\left(\hat{Q}_\theta(Y|\boldsymbol{x}_k)\right) \tag{4.6.5}$$

La (4.6.5) implica che i quantili condizionati della variabile risposta trasformata sono equivalenti ai quantili condizionati trasformati della variabile risposta. Solo se $g()$ è affine, la media condizionata assume la medesima proprietà, nel rimanente insieme di casi no, data la disuguaglianza di Jensen.

L'ultima proprietà è definita come invarianza dei quantili per trasformazioni monotone e fornirà un contributo decisivo nell'impostazione del modello tariffario introdotto in questa tesi nei paragrafi successivi.

Il tutto deriva dal fatto che se $g()$ è monotona:

$$P(Y \leq y) = P\big(g(Y) \leq g(y)\big)$$

L'uguaglianza (4.6.5) può essere molto utile in determinate condizioni, ad esempio se la risposta viene trasformata per il logaritmo, si è sempre perfettamente giustificati nell'interpretare $e^{\boldsymbol{x}_k\beta_\theta}$ come una stima appropriata del quantile condizionato di livello θ della Y; cosa non giustificabile formalmente per la media condizionata (Koenker, 2005).

Tra tutte le caratteristiche della Quantile Regression, una merita sicuramente particolare attenzione: la scarsa sensibilità delle stime dei quantili condizionati, rispetto a valori che possono essere considerati outliers (Koenker 2005). Questo significa che qualsiasi osservazione, anche estrema, della variabile risposta non altera, nella stragrande maggioranza dei casi, la stima dei quantili condizionati che produce il modello di regressione. La stessa cosa non vale per il modello lineare. Infine una delle ultime considerazioni può riguardare l'efficienza. Come già introdotto all'inizio del capitolo, per una vasta gamma di distribuzioni non normali della variabile d'interesse, la varianza asintotica della mediana campionaria è all'incirca il 50% più piccola, se non ancora nettamente più piccola, della varianza asintotica della media campionaria; solo nel caso di una distribuzione normale il rapporto si inverte e la media risulta più efficiente. Koenker e Bassett (1978) estendono questo concetto legato a misure campionarie a quello della regressione del quantile, affermando che la mediana condizionata è uno stimatore più efficiente di quello ottenuto con i minimi quadrati ordinari sotto ipotesi distributive non gaussiane. È quindi ragionevole sacrificare l'ipotesi di normalità e utilizzare la regressione del quantile per ottenere un notevole miglioramento, rispetto ai minimi quadrati ordinari, in termini di efficienza. Quest'ultima, inoltre, può essere aumentata con opportune procedure di ponderazione qualora le osservazioni presentino strutture di correlazione e eteroschedasticità (Koenker, 2010b).

L'interpretazione dei parametri nella regressione del quantile è simile a quella dei parametri del modello di regressione lineare classico. In quest'ultimo, infatti, per la $j-$esima variabile il parametro associato è β_j, che altro non è che la derivata parziale del valore atteso condizionato di y dato $\boldsymbol{x}_k$ (ceteris paribus):

$$\beta_j = \frac{dE(Y|X)}{d\boldsymbol{x}_{kj}},$$

da interpretarsi come l'effetto marginale che la $j-$esima variabile ha sulla media condizionata. Per il modello di regressione del quantile l'interpretazione è analoga a quella della regressione lineare, a meno del

fatto che gli effetti marginali di una variabile (ceteris paribus) sono da riferirsi al quantile condizionato di livello θ. Se il quantile condizionato di livello θ viene definito come $Quant_\theta(Y|\boldsymbol{x}_k) = \sum_{j=1}^{m} \beta_{\theta j} x_{kj}$, l'effetto marginale della j-esima variabile sul quantile condizionato di livello θ è:

$$\frac{dQuant_\theta(Y|\boldsymbol{x}_k)}{d\boldsymbol{x}_{kj}} = \beta_{\theta k}.$$

Siccome la regressione del quantile gode della proprietà di equivarianza rispetto a trasformazioni monotone della risposta (equazione (4.6.5)), se si sta eseguendo una regressione del quantile su trasformazioni logaritmiche, allora possiamo definire l'effetto marginale della j-esima variabile sul quantile di livello θ della variabile risposta y come:

$$\frac{dQuant_\theta(Y|\boldsymbol{x}_k)}{d\boldsymbol{x}_{kj}} = \beta_{\theta k} e^{\boldsymbol{x}_{kj}\beta_\theta} .$$

5. IL GLM E LA QUANTILE REGRESSION NEL CONTESTO DELLA FIE

5.1. I LIMITI DEL GLM

5.1.1. L'IMPOSSIBILITÀ DI COPRIRE IL FABBISOGNO PURO

Riprendendo la notazione del paragrafo 2.5, dalle equazioni (3.3.6) e (3.4.2) si evince che i modelli lineari generalizzati forniscono, in prima analisi, una stima del valore atteso di Y condizionato alle caratteristiche dell'individuo definite dal vettore $\boldsymbol{x}_{ij}$.

Formalmente si avrà che:

$$P_{ij} = P^{(0)} \cdot \beta_{1i}^{GLM} \cdot \beta_{2j}^{GLM} = E(Y|\boldsymbol{x}_{ij}) \tag{5.1.1}$$

Per cui sostituendo la (5.1.1)**,** nell'equazione (2.5.1) si ottiene che la stima GLM, non permette di coprire interamente il fabbisogno puro $Quant_\theta[\ddot{Y}]$:

$$Quant_\theta[\ddot{Y}] \neq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij}E(Y|\boldsymbol{x}_{ij}) \tag{5.1.2}$$

Risulta, infatti, che $Quant_\theta[\ddot{Y}] > \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij}E(Y|\boldsymbol{x}_{ij})$.

Tale risultato è dimostrato applicando il valore atteso ad ambo i membri della (5.1.2)

$$E\left[Quant_\theta[\ddot{Y}]\right] \neq E\left[\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij}E(Y|\boldsymbol{x}_{ij})\right] \leftrightarrow$$

$$Quant_\theta[\ddot{Y}] \neq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij}E\left[E(Y|\boldsymbol{x}_{ij})\right] \leftrightarrow$$

$$Quant_\theta[\ddot{Y}] \neq E[Y]\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij} \leftrightarrow$$

$$Quant_\theta[\ddot{Y}] \neq \mathrm{r} \cdot E[Y] = E[\ddot{Y}] \tag{5.1.3}$$

Poiché:

$$Quant_\theta[\ddot{Y}] = E[\ddot{Y}] + m[\ddot{Y}],$$

risulta evidente:

$$Quant_\theta[\ddot{Y}] > \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij}E(Y|\boldsymbol{x}_{ij}).$$

I GLM permettono solo un'allocazione della componente equa del fabbisogno, e non permettono la copertura della componente di caricamento del fabbisogno puro $m[\ddot{Y}] = Quant_\theta[\ddot{Y}] - E[\ddot{Y}]$, in funzione delle caratteristiche individuali.

Nella pratica, per ottenere l'equilibrio attuariale, viene calcolato un coefficiente correttivo moltiplicativo $\tilde{P}^{(0)}$, per sanare il disequilibrio indotto dalla stima GLM. Tale coefficiente, osservando la (5.1.3) risulta pari a:

$$\tilde{P}^{(0)} = \frac{Quant_{\theta}[\ddot{Y}]}{E[\ddot{Y}]}; \tag{5.1.4}$$

di conseguenza il premio stimato con i GLM, è pari a:

$$P_{ij} = \tilde{P}^{(0)} \cdot E\big(Y|\boldsymbol{x}_{ij}\big).$$

Dall'osservazione della (5.1.4), si può concludere immediatamente che il fattore di correzione $\tilde{P}^{(0)}$, identico per tutte le classi di rischio, non dipende dalle caratteristiche degli assicurati. In sintesi, attraverso tale approccio il premio puro per profilo di rischio è riconducibile al principio di calcolo del valore atteso, secondo la seguente relazione:

$$P_{ij} = E\big[Y|\boldsymbol{x}_{ij}\big] + \gamma E\big[Y|\boldsymbol{x}_{ij}\big] = E\big[Y|\boldsymbol{x}_{ij}\big] \cdot (1+\gamma)$$

poiché

$$P_{ij} = \tilde{P}^{(0)} \cdot E\big(Y|\boldsymbol{x}_{ij}\big) = E\big(Y|\boldsymbol{x}_{ij}\big) \cdot \left(\frac{Quant_{\theta}[\ddot{Y}]}{E[\ddot{Y}]}\right)$$

Allora:

$$1+\gamma = \frac{Quant_{\theta}[\ddot{Y}]}{E[\ddot{Y}]} \leftrightarrow$$

$$\gamma = \frac{Quant_{\theta}[\ddot{Y}] - E[\ddot{Y}]}{E[\ddot{Y}]} = \frac{m[\ddot{Y}]}{E[\ddot{Y}]}$$

Si evince, dunque, che la procedura adottata, definisce un caricamento di sicurezza per ogni profilo $m\big[Y_{ij}\big] = \gamma E\big[Y|\boldsymbol{x}_{ij}\big]$, ma il caricamento stesso è proporzionale al valore atteso attraverso un coefficiente unico indipendente dalle caratteristiche di rischiosità specifiche del profilo.

Dalla relazione (5.1.3), si evince che tale approccio non permette una allocazione della componente di caricamento del fabbisogno puro $m\big[\ddot{Y}\big]$, dipendente dalla distribuzione di probabilità per singolo profilo di rischio, ma esclusivamente del suo momento primo.

In tal modo due profili con stessa media, ma con diverse caratteristiche in termini di dispersione, pagherebbero lo stesso premio.

Una prima proposta originale per risolvere tale limite, può essere quella di sfruttare i GLM che oltre alla stima della componente attesa condizionata, forniscono anche una stima della distribuzione di probabilità condizionata della variabile risposta. Tale importante caratteristica, non sfruttata nella pratica attuariale, permette di individuare misure di sintesi come i quantili, che potrebbero essere funzionali ad una copertura

immediata del fabbisogno puro, senza dover fare ricorso al fattore correttivo $\tilde{P}^{(0)}$. Tuttavia anche tale soluzione, come dimostrato nel paragrafo successivo, non risulta pienamente efficiente.

L'introduzione della Quantile Regression in questo lavoro ha proprio la funzione di correggere tali lacune, individuando una stima diretta del quantile condizionato e quindi direttamente del premio puro.

5.1.2. L'INADEGUATEZZA DEI GLM NELLA STIMA DEI QUANTILI CONDIZIONATI PER PROFILO..

Facendo riferimento al paragrafo (3.4.1), si ricorda che i GLM nella pratica vengono utilizzati sfruttando la fattorizzazione del valore atteso della quota danni in frequenza media e costo medio:

$E[Y] = E[N] \cdot E[Z]$.

L'esecuzione dei GLM su N e Z, permette la definizione della distribuzione di probabilità per profilo sia della frequenza che del costo medio. Risulta chiaro che se la variabile N, in rami come l'RCA, ha una variabilità abbastanza contenuta, la variabile Z è molto più volatile e per i nostri obiettivi di stima di un caricamento per profilo, risulta quella di maggior interesse.

Soffermando l'attenzione proprio su Z, l'ipotesi distributiva maggiormente utilizzata è di tipo Gamma e, come noto in letteratura (Gigante, Picech, Sigalotti 1998), la stima GLM viene eseguita in questi casi ipotizzando un parametro di forma costante. Formalmente:

$Z_{ij} = Z|\boldsymbol{x}_{ij} \sim Gamma(a, b_{ij})$

Siano:

- $F_{Z_{ij}}(v) = \int_0^v \frac{1}{\Gamma(a) b_{ij}^a} x^{a-1} e^{-\frac{x}{b_{ij}}} dx$
- $E[Z_{ij}] = a \cdot b_{ij}$
- $VAR[Z_{ij}] = a \cdot b_{ij}^2$

Per cui, se il parametro di forma è costante, tutte le distribuzioni di probabilità per profilo hanno lo stesso coefficiente di variazione:

$$\frac{E[Z_{ij}]}{\sqrt{VAR[Z_{ij}]}} = \frac{1}{\sqrt{a}}$$

Di conseguenza, nel caso della Gamma, è possibile definire il caricamento di sicurezza specifico di ogni profilo a partire dalla distribuzione di probabilità associata allo stesso. Tuttavia, a causa dell'ipotesi di coefficiente di variazione costante, è possibile dimostrare che tale caricamento risulta, come nel caso precedente, proporzionale al valore atteso. La dimostrazione è riportata nel seguente lemma.

Lemma:

Date due variabili casuali Gamma Z_{ij} e Z_{nm}, con lo stesso parametro di forma a e diverso parametro di scala b, i quantili a livello di probabilità θ possono essere espressi come funzione dei valori attesi $E[Z_{ij}]$ e $E[Z_{nm}]$, attraverso un'unica costante di proporzionalità $\rho \in \mathbb{R}$.

IPOTESI

$Z_{ij} \sim Gamma(a, b_{ij})$;

$Z_{nm} \sim Gamma(a, b_{nm})$;

TESI:

$$\frac{Quant_{\theta}[Z_{ij}]}{E[Z_{ij}]} = \frac{Quant_{\theta}[Z_{nm}]}{E[Z_{nm}]} = \rho$$

DIMOSTRAZIONE:

Poiché:

$$\theta = Prob\{Z_{ij} < Quant_{\theta}[Z_{ij}]\}$$

La tesi è dimostrata se:

$$Prob\{Z_{ij} < Quant_{\theta}[Z_{ij}]\} = \int_0^{\rho \cdot a \cdot b_{ij}} \frac{1}{\Gamma(a) b_{ij}^a} x^{a-1} e^{-\frac{x}{b_{ij}}} dx = \int_0^{\rho \cdot a \cdot b_{nm}} \frac{1}{\Gamma(a) b_{nm}^a} x^{a-1} e^{-\frac{x}{b_{nm}}} dx = Prob\{Z_{nm} < Quant_{\theta}[Z_{nm}]\}.$$

$$\int_0^{\rho \cdot a \cdot b_{ij}} \frac{1}{\Gamma(a) b_{ij}^a} x^{a-1} e^{-\frac{x}{b_{ij}}} dx = \int_0^{\rho \cdot a \cdot b_{nm}} \frac{1}{\Gamma(a) b_{nm}^a} x^{a-1} e^{-\frac{x}{b_{nm}}} dx \leftrightarrow$$

Si elimina il termine $\Gamma(a)$ e si moltiplicano ambo i membri per $b_{ij}^a * b_{nm}^a$

$$b_{nm}^a \int_0^{\rho \cdot a \cdot b_{ij}} x^{a-1} e^{-\frac{x}{b_{ij}}} dx = b_{ij}^a \int_0^{\rho \cdot a \cdot b_{nm}} x^{a-1} e^{-\frac{x}{b_{nm}}} dx \leftrightarrow$$

Si derivano ambo i membri rispetto a ρ:

$$\frac{d}{d\rho}\left[b_{nm}^a \int_0^{\rho \cdot a \cdot b_{ij}} x^{a-1} e^{-\frac{x}{b_{ij}}} dx\right] = \frac{d}{d\rho}\left[b_{ij}^a \int_0^{\rho \cdot a \cdot b_{nm}} x^{a-1} e^{-\frac{x}{b_{nm}}} dx\right] \leftrightarrow$$

$$a \cdot b_{ij} \cdot b_{nm}^a \cdot (\rho \cdot a \cdot b_{ij})^{a-1} e^{-\rho * a} = a \cdot b_{ij}^a \cdot b_{nm} (\rho \cdot a \cdot b_{nm})^{a-1} \cdot e^{-\rho * a} \leftrightarrow$$

$$a^a \cdot b_{ij}^a \cdot b_{nm}^a \cdot \rho^{a-1} \cdot e^{-\rho * a} = a^a \cdot b_{ij}^a \cdot b_{nm}^a \cdot \rho^{a-1} \cdot e^{-\rho * a}$$

Risulta in tal modo mostrata la tesi.

Stimare, quindi, le distribuzioni di probabilità Z_{ij}, attraverso il GLM gamma, equivarrebbe di fatto ad assegnare la medesima variabilità a tutti i profili, e i quantili di tutte le distribuzioni per profilo avrebbero un rapporto costante con i rispettivi valori attesi.

Tale risultato presuppone un'ipotesi molto forte dei GLM gamma, e molto spesso irrealistica, nella stima della distribuzione di probabilità per profilo, cioè che al variare della classe di rischio il parametro di forma della distribuzione resti sempre lo stesso mentre varia solo quello di scala.

Tale ipotesi fa sì che il GLM sia, nella grande maggioranza dei casi inadeguato nella stima dei quantili per profilo, come si mostrerà nell'applicazione.

I punti fondamentali che ci inducono a introdurre una tecnica di tariffazione alternativa al GLM sono dunque:

- mancata allocazione della componente $Quant_\theta[\ddot{Y}] - E[\ddot{Y}]$ dovuta all'uso che si fa nella pratica dei GLM, ossia la definizione della sola stima del valore atteso condizionato della variabile risposta;
- l'uso del GLM come tecnica per la stima della distribuzione di probabilità per profilo, è inadeguato, per le ipotesi sottostanti le stime dei parametri.

5.2. LA DEFINIZIONE DI UN MODELLO QUANTILE REGRESSION PER LA TARIFFAZIONE.

Una volta introdotta la Quantile Regression da un punto di vista teorico e meramente statistico si definisce, in questo paragrafo, l'idea per inserirla nel contesto della tariffazione. A oggi in ambito attuariale la QR è stata utlizzata per il pricing di contratti contro il furto in un lavoro di Kudryavtsev (2009), ma in modi totalmente diversi rispetto all'impianto modellistico pensato per questo lavoro. Per comprendere meglio la ratio alla base del modello di pricing introdotto nella tesi si deve pensare che il passaggio dal GLM alla QR comporta la necessità di operare con quantili condizionati, quando prima si operava con medie condizionate ed è chiaro che questo implica, oltre ai vantaggi già descritti nei precedenti capitoli, la perdita di alcune proprietà della media particolarmente vantaggiose ai fini tariffari. L'introduzione della Quantile Regression nel contesto della personalizzazione del premio presenta, dunque, alcune criticità.

Considerando la (5.1.1), adattata alla nostra collettività di riferimento di r rischi assicurati, si evince che la Quantile Regression fornisce, in prima analisi, una stima del quantile di Y, condizionato alle caratteristiche dell'individuo, a un certo livello di probabilità θ.

Formalmente si avrà che:

$$P_{ij} = P^{(0)} * \beta_{\theta 1i} * \beta_{\theta 2j} = Quant_\theta(Y|\boldsymbol{x}_{ij}) \qquad (5.2.1)$$

Tuttavia, occorre constatare che la natura semicontinua della variabile casuale Y introduce dei problemi, (come spiegato nel paragrafo (4.4)), sia nel contesto dell'unicità della soluzione di minimo, ma anche in quello di stima della sparsity function e quindi di definizione della stima della matrice di covarianza

asintotica e di alcuni tipi di intervalli di confidenza. L'idea quindi è di superare gli inconvenienti generati dalla componente discreta di Y, costruendo un impianto tariffario che utilizzi la Quantile Regression solo sulla componente di Y a realizzazioni strettamente positive.

Facendo riferimento al caso GLM, si ricorda che nella pratica attuariale si considera la fattorizzazione del valore atteso di Y:

$E[Y] = E[N]E[Z]$,

quindi anche in questo caso non si effettua un'analisi sulla variabile aleatoria semicontinua Y, ma sulle due componenti indipendenti numero dei sinistri e importo del danno associato ad un sinistro.

Si fa notare che l'estensione della medesima soluzione al caso Quantile Regression, non risolverebbe il problema dovuto alla semicontinuità di Y, in quanto anche in condizioni di indipendenza tra N e Z:

$Q_\theta[Y] \neq Q_\theta[N]Q_\theta[Z]$.

La soluzione proposta in questo lavoro è quella di introdurre un modello a due parti.

5.3. IL MODELLO A DUE PARTI E LA DISTRIBUZIONE LAPLACE ASIMMETRICA

Per introdurre il modello a due parti si considerano le seguenti grandezze:

- $I_{\{y=0\}}$ l'indicatore dell'evento "l'assicurato non è colpito da sinistro",
- $I_{\{y>0\}}$ l'indicatore dell'evento "l'assicurato è colpito da almeno un sinistro",
- $P(Y > 0) = P\big(I_{\{y>0\}} = 1\big)$ la probabilità che l'assicurato sia colpito da almeno un sinistro,

si può esprimere la funzione di densità della variabile Y nel seguente modo:

$$f_Y(y) = \big(1 - P(Y > 0)\big) \cdot I_{\{y=0\}} + P(Y > 0) \cdot f_{Y|Y>0}(y), \qquad (5.3.1)$$

dove con $f_{Y|Y>0}$ si indica la densità di Y condizionata all'avvenimento di almeno un sinistro.

La funzione di densità di Y, scritta nella forma (5.3.1), si dirà associata ad un modello a due parti di tipo semicontinuo. Si vuole passare, quindi, da un modello tariffario che sfrutti la fattorizzazione del valore atteso ad uno che sfrutti la decomposizione della funzione di densità di Y, attraverso un modello a due parti.

Supponendo di avere a disposizione un campione di n realizzazioni della variabile Y, e il relativo set di covariate, è possibile generalizzare la (5.3.1) al caso regressivo. Infatti se si introduce:

- il vettore riga di covariate $\boldsymbol{x}_k = (x_{k1}, \dots\dots\dots, x_{km})$,
- il set di parametri δ, utilizzato per personalizzare la probabilità di effettuare sinistro,
- il set di parametri ρ, utilizzato per personalizzare il costo non nullo per assicurato,

- $p_{\delta,\boldsymbol{x}_k} = P(Y > 0)|\boldsymbol{x}_k$ la probabilità di effettuare almeno un sinistro condizionata alle caratteristiche dell'assicurato,
- $g_{\rho,\boldsymbol{x}_k}(y_k) = f_{Y|[(y_k>0)\cap \boldsymbol{x}_k]}(y_k)$, la densità condizionata a valori positivi di Y,

è possibile definire la funzione di verosimiglianza associata al modello (5.3.1):

$$L(\boldsymbol{\delta}, \boldsymbol{\rho}) = \prod_{\mathrm{k}} f_Y(y_k) = \prod_{y_k=0}\left(1 - p_{\delta,\boldsymbol{x}_k}\right) \prod_{y_k>0} p_{\delta,\boldsymbol{x}_k}\, g_{\rho,\boldsymbol{x}_k}(y_k). \tag{5.3.2}$$

Duan (1983) ha mostrato che qualunque sia il modello scelto per stimare p_{δ,x_k} e g_{ρ,x_k}, è possibile fattorizzare la verosimiglianza (5.3.2) in due parti: una relativa soltanto ai parametri $\boldsymbol{\delta}$ e l'altra relativa ai parametri $\boldsymbol{\rho}$.

$$L(\boldsymbol{\delta}, \boldsymbol{\rho}) = L_1(\boldsymbol{\delta}) L_2(\boldsymbol{\rho}) = \prod_{y_k=0}\left(1 - p_{\delta,\boldsymbol{x}_k}\right) \cdot p_{\delta,\boldsymbol{x}_k} \prod_{y_k>0} g_{\rho,\boldsymbol{x}_k}(y_k).$$

Questa fattorizzazione permette di massimizzare $L_1(\boldsymbol{\delta})$ ed $L_2(\boldsymbol{\rho})$ separatamente.

La nuova tecnica che si vuole sperimentare per calcolare il premio, consiste nel modellare la probabilità di effettuare almeno un sinistro con un GLM di tipo Logit o Probit e solo in un secondo momento nell' effettuare una Quantile Regression, sulla parte di campione con determinazioni strettamente positive, ovvero sulla variabile casuale:

$(Y|\boldsymbol{y} > 0) = \dot{Y}$.

La struttura del modello a due parti richiede tuttavia la presenza di una funzione di densità che rappresenti la parte di campione sinistrata, ovvero $g_{\rho,\boldsymbol{x}_k}(y_k)$. Se si vuole applicare la Quantile Regression alla parte di campione con determinazioni positive, occorre definire una funzione di densità associabile al modello QR.

Si può dimostrare che risolvere il problema di minimo (4.3.1), equivale a massimizzare la funzione di verosimiglianza associata a una funzione di densità di tipo Laplace asimmetrica (di seguito **ALD**).

Si consideri il problema di minimo (4.3.1), sulla variabile $\dot{Y}$, che riportiamo per comodità:

$$min_{\beta_\theta}\left\{\sum_{k:y_k\geq \boldsymbol{x}_k'\beta_\theta} \theta|y_k - \boldsymbol{x}_k'\beta_\theta| + \sum_{k:y_k<\boldsymbol{x}_k'\beta_\theta}(1-\theta)|y_k - \boldsymbol{x}_k'\beta_\theta|\right\}$$

Si definisca la ALD:

$$f(y, x\beta_\theta, \sigma) = \frac{\theta(1-\theta)}{\sigma} exp\left\{-\rho_\theta\left(\frac{y-\boldsymbol{x}'\beta_\theta}{\sigma}\right)\right\}, y \in (-\infty, +\infty), \tag{5.3.3}$$

dove $\rho_\theta(u)$ è definito dalla (4.2.2).

A questo punto si supponga di disporre di un campione di numerosità n della variabile $\dot{Y}$, dove $\dot{Y}$ è Laplace Asimmetrica (di seguito $\dot{Y}\sim\boldsymbol{AL}$**)**. La funzione di verosimiglianza è:

$$L(\boldsymbol{y}, \boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma) = \prod_{k=1}^{n} f(y_k, \boldsymbol{x}_k'\beta_\theta, \sigma) = \frac{\theta^n(1-\theta)^n}{\sigma^n} exp\left\{-\sum_{k=1}^{n} \rho_\theta\left(\frac{y_k - \boldsymbol{x}_k'\beta_\theta}{\sigma}\right)\right\}.$$

Supponendo σ e θ noti, possiamo eliminare il fattore $\frac{\theta^n(1-\theta)^n}{\sigma^n}$, quindi:

$$L(\boldsymbol{y}, \boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma) = \ exp\left\{-\sum_{k=1}^{n} \rho_\theta\left(\frac{y_k - \boldsymbol{x}_k'\beta_\theta}{\sigma}\right)\right\}.$$

Poichè il logaritmo è una funzione monotona crescente, i punti di massimo di $L(\boldsymbol{y}, \boldsymbol{x}\boldsymbol{\beta_\theta}, \sigma)$ equivalgono a quelli di

$$l(\boldsymbol{y}, \boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma) = log\big(L(\boldsymbol{y}, \boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma)\big) = \left\{-\sum_{k=1}^{n} \rho_\theta\left(\frac{y_k - \boldsymbol{x}_k'\beta_\theta}{\sigma}\right)\right\}.$$

Per ottenere gli stimatori di massima verosimiglianza, si deve risolvere il problema:

$$max_{\boldsymbol{\beta_\theta}}\big(l(\boldsymbol{y}, \boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma)\big),$$

il che equivale a risolvere:

$$min_{\boldsymbol{\beta_\theta}}\big(-l(\boldsymbol{y}, \boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma)\big) \ = min_{\beta_\theta}\left\{\sum_{k:y_k \geq \boldsymbol{x}_k'\beta_\theta} \theta \left|\frac{y_k - \boldsymbol{x}_k'\beta_\theta}{\sigma}\right| + \sum_{k:y_k < \boldsymbol{x}_k'\beta_\theta} (1-\theta)\left|\frac{y_k - \boldsymbol{x}_k'\beta_\theta}{\sigma}\right|\right\}$$

e poiché σ è noto:

$$min_{\boldsymbol{\beta_\theta}}\big(-l(\boldsymbol{y}, \boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma)\big) = min_{\beta_\theta}\left\{\sum_{k:y_k \geq \boldsymbol{x}_k'\beta_\theta} \theta|y_k - \boldsymbol{x}_k'\beta_\theta| + \sum_{k:y_k < \boldsymbol{x}_k'\beta_\theta} (1-\theta)|y_k - \boldsymbol{x}_k'\beta_\theta|\right\}.$$

Si deduce, quindi, che stimare i parametri di una Quantile Regression minimizzando la (4.3.1) equivale a trovare gli stimatori di massima verosimiglianza se la variabile casuale $\dot{Y}$ si distribuisce come una Laplace asimmetrica di parametri $(\boldsymbol{x}'\boldsymbol{\beta_\theta}, \sigma)$.

Si nota, dalla (5.3.3), che la ALD definisce la densità di una variabile casuale avente come supporto tutto l'asse reale. Poiché la variabile casuale $\dot{Y}$ ha realizzazioni esclusivamente positive, non si può eseguire una Quantile Regression direttamente su $\dot{Y}$.

L'idea è quella di definire una Quantile Regression sui logaritmi degli importi, per cui si ipotizza:

$V = \log(\dot{Y}) \sim AL(x\beta_\theta, \sigma)$.

Sarà sufficiente, allora, eseguire la Quantile Regression sui logaritmi degli importi del danno (che è una Quantile Regression canonica, poiché $V = \log(\dot{Y}) \sim AL(x\beta_\theta, \sigma)$), quindi relativamente a grandezze che possono assumere anche valori negativi. In tal modo l'output della Quantile Regression è $Q_\theta(\log(\dot{Y}|x_k))$.

Sfruttando la proprietà di invarianza del quantile rispetto a trasformazioni monotone, definita dalla (4.6.5) che per comodità è di seguito riportata:

$$Q_\theta(\log(\dot{Y}|x_k)) = \log(Q_\theta(\dot{Y}|x_k)),$$

eseguendo l'esponenziale degli output ottenuti con la QR, ovvero di $Q_\theta(\log(\dot{Y}|x_k))$, si ottiene:

$$e^{Q_\theta(\log(\dot{Y}|x_k))} = e^{\log(Q_\theta(\dot{Y}|x_k))} = Q_\theta(\dot{Y}|x_k),$$

dove la prima uguaglianza è spiegata dalla (4.6.5)

In sintesi gli step per stimare i parametri del modello a due parti sono i seguenti:

- Esecuzione di un modello logit per stimare la verosimiglianza associata alla componente discreta del modello a due parti
- Esecuzione di un modello QR per stimare la verosimiglianza associata alla componente continua del modello a due parti.

Il fatto che la seconda componente del modello a due parti sia una Quantile Regression, permette la definizione non solo della componente equa del premio individuale (come accade nel caso GLM), ma anche della componente di caricamento individuale. In sintesi il premio, output di tale procedura, può essere ancora definito come un quantile, per cui: $P_{ij} = Quant_\theta(Y|\boldsymbol{x}_{ij})$.

5.4. L'EFFETTO DIVERSIFICAZIONE E IL LIVELLO DI PROBABILITÀ OTTIMO.

Una volta superato il problema relativo alla semicontinuità di Y, con l'introduzione del modello a due parti, occorre dire che se quest'ultimo fosse calibrato per definire il quantile condizionato di Y allo stesso livello di probabilità del fabbisogno puro, la somma dei premi così calcolati non realizzerebbe la (2.6.2), poiché non è garantita la proprietà di additività dei quantili. Ricordando che il quantile è di fatto una misura di rischio di tipo VaR, come noto può non rispettare le condizioni di additività e soprattutto quelle di sub-additività, quest'ultima implica la mancanza di coerenza nel senso di Artzner (1999). La differenza tra una misura di rischio applicata alla somma di variabili aleatorie e la somma della misura di rischio applicata alle singole variabili aleatorie è definita come "effetto diversificazione".

Infatti poiché il modello a due parti definisce:

$$P_{ij} = Quant_{\theta}(Y|\boldsymbol{x}_{ij})$$

inserendo tale relazione, nella (2.6.2) si ottiene:

$$Quant_{\theta}[\ddot{Y}] \neq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij} Quant_{\theta}(Y|\boldsymbol{x}_{ij})$$

Tuttavia, la soluzione proposta in questo lavoro è quella di determinare il livello di probabilità θ^* tale che:

$$Quant_{\theta}[\ddot{Y}] = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij} Quant_{\theta^*}(Y|\boldsymbol{x}_{ij}) .$$

A tal fine si è definito un algoritmo che, con una certa tolleranza, (il più possibile piccola) permetta l'individuazione di θ^*.

Indicando con ε la tolleranza definiamo l'algoritmo con i seguenti step:

1. Si inizializza un contatore $h = 1$
2. Si fissano $\theta_{(1)}^{left} = 0$, $\theta_{(1)}^{right} = 1$ e $\theta_{(1)} = 0{,}5$
3. Si effettua una Quantile Regression per $\theta = \theta_{(h)}$
4. Si calcola $Quant_{\theta}[\ddot{Y}] - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij} Quant_{\theta_{(h)}}(Y|\boldsymbol{x}_{ij})$
   a. Se $\left|Quant_{\theta}[\ddot{Y}] - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij} Quant_{\theta_{(h)}}(Y|\boldsymbol{x}_{ij})\right| \leq \varepsilon$
      - $\theta^* = \theta_{(h)}$
   b. Se $Quant_{\theta}[\ddot{Y}] - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij} Quant_{\theta_{(h)}}(Y|\boldsymbol{x}_{ij}) > \varepsilon$
      - $\theta_{(h+1)} = \frac{\theta_{(h)} + \theta_{(h)}^{right}}{2}$,
      - $\theta_{(h+1)}^{left} = \theta_{(h)}$
      - $\theta_{(h+1)}^{right} = \theta_{(h)}^{right}$
      - $h = h + 1$
      - torno al punto 3
   c. Se $Quant_{\theta}[\ddot{Y}] - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij} Quant_{\theta_{(h)}}(Y|\boldsymbol{x}_{ij}) < -\varepsilon$
      - $\theta_{(h+1)} = \frac{\theta_{(h)} + \theta_{(h)}^{left}}{2}$,
      - $\theta_{(h+1)}^{right} = \theta_{(h)}$
      - $\theta_{(h+1)}^{left} = \theta_{(h)}^{left}$
      - $h = h + 1$
      - torno al punto 3

Utilizzando questa procedura si ottiene un premio per profilo funzione delle caratteristiche del singolo individuo, che consente di allocare il fabbisogno puro ed in particolare la componente di caricamento $m[\ddot{Y}]$ sulle singole teste assicurate. In linea di principio il livello di probabilità θ^*, che individua il margine di sicurezza da applicare al singolo profilo, dovrebbe essere inferiore al livello di probabilità complessivo θ; tuttavia il VaR è una misura di rischio che non garantisce tale proprietà , poiché non rispetta la proprietà di sub-additività.

6. APPLICAZIONE.

Per dimostrare le differenze tra il nuovo approccio proposto e quanto eseguito tradizionalmente nella pratica attuariale attraverso l'impiego dei GLM, nel seguito si riportano alcune risultanze delle analisi numeriche condotte. Prima di calcolare il premio tramite modello a due parti, si sono effettuati degli esperimenti su diversi database, in modo tale da analizzare le caratteristiche della Quantile Regression sotto diversi punti di vista. Successivamente è stato calcolato il premio tramite modello a due parti su un dataset reale preso dall'esperienza di una compagnia danni italiana operante nel ramo R.C.A..

6.1. CONFRONTO TRA IL MODELLO GAMMA E LA QUANTILE REGRESSION

Nei primi esempi si sono generati dei campioni fittizi esclusivamente sulla variabile (Z) importo del danno associato ad un sinistro (variabile continua), per effettuare dei confronti tra la QR e il GLM gamma. Nel primo di questi casi si è definito un campione di importi da un'unica variabile casuale gamma $\Gamma(a, b)$ con gli stessi parametri per tutti i rischi in portafoglio. Tale campione è stato ripartito in sei classi di rischio, relative ai seguenti fattori di rischio:

- il sesso con modalità Maschio e Femmina;
- l'età, che è stata considerata qualitativa con modalità Alta, Media e Bassa.

Il campione presenta in totale 3000 valori e i 6 profili di rischio hanno diversa numerosità in base a quanto definito in Tabella 2

Tabella 2: distribuzione di frequenza sesso età

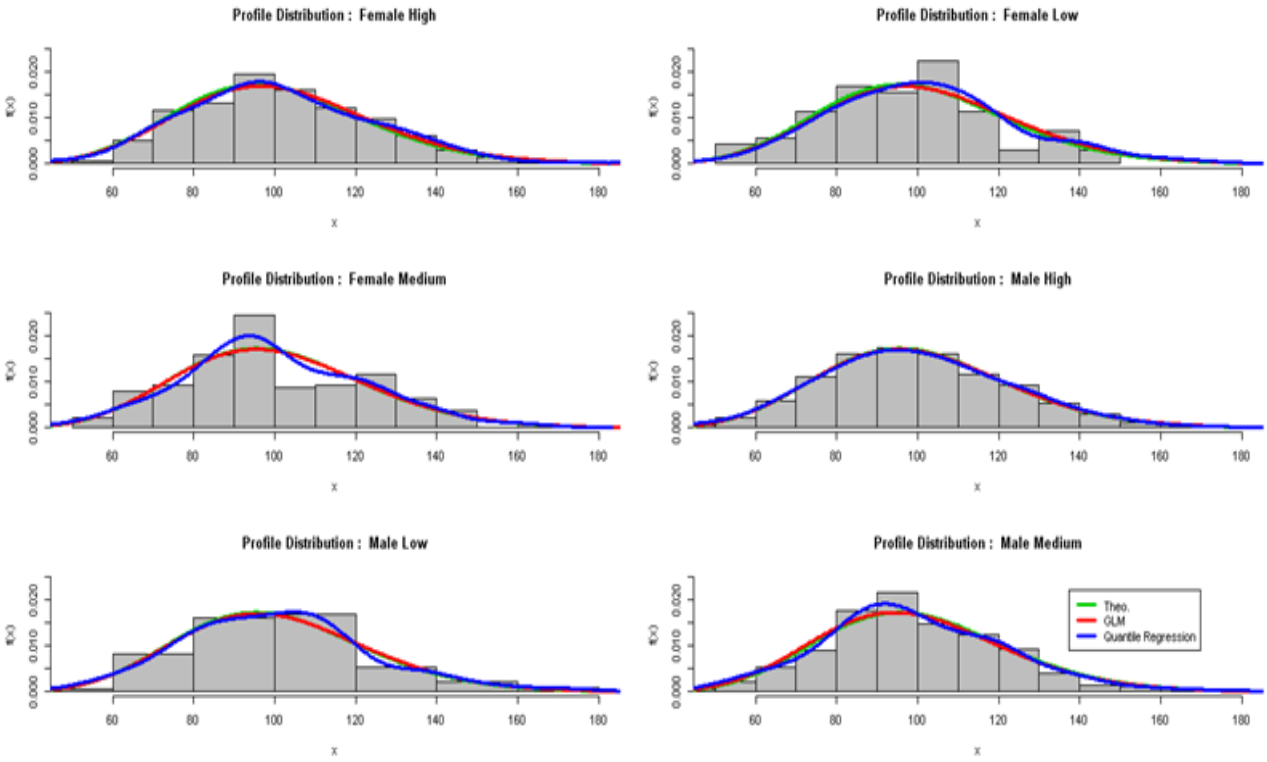
	A	B	M	Totale complessivo
F	527	73	159	**759**
M	1582	209	450	**2241**
Totale complessivo	**2109**	**282**	**609**	**3000**

In Figura 11 sono riportati in ciascuna delle sezioni gli istogrammi della distribuzione campionaria, per ogni profilo di rischio. Si notano delle differenze tra gli istogrammi, nonostante il campione sia generato sempre dalla stessa distribuzione, a causa di fluttuazioni casuali. Sempre riferendosi a un singolo profilo di rischio:

- la curva verde rappresenta la distribuzione teorica da cui sono stati generati i dati ed ovviamente è la stessa per ogni profilo,
- la linea rossa rappresenta la funzione di densità ottenuta stimando i parametri attraverso un GLM di tipo gamma. $\Gamma(\hat{a}, \hat{b}_{ij})$
- la linea blu rappresenta la densità definita attraverso una Quantile Regression. Essa è stata ottenuta effettuando QR a diversi livelli di probabilità (livelli da 0 a 1 con passo 0,00125, quindi circa 800); i valori fittati dei modelli sono stati utilizzati come input di uno stimatore Kernel per densità.

Si può affermare, dall'osservazione del grafico, che in questo primo caso sia il GLM che la Quantile Regression presentano un fitting abbastanza soddisfacente e abbastanza simile. Si fa notare, tuttavia, che i dati in questione sono stati generati da una distribuzione gamma uguale per tutti i profili e che l'ipotesi distributiva per la variabile risposta del GLM è anch'essa di tipo gamma.

Figura 11: Distribuzioni di probabilità per profilo, in cui il campione di partenza è generato da un'unica gamma



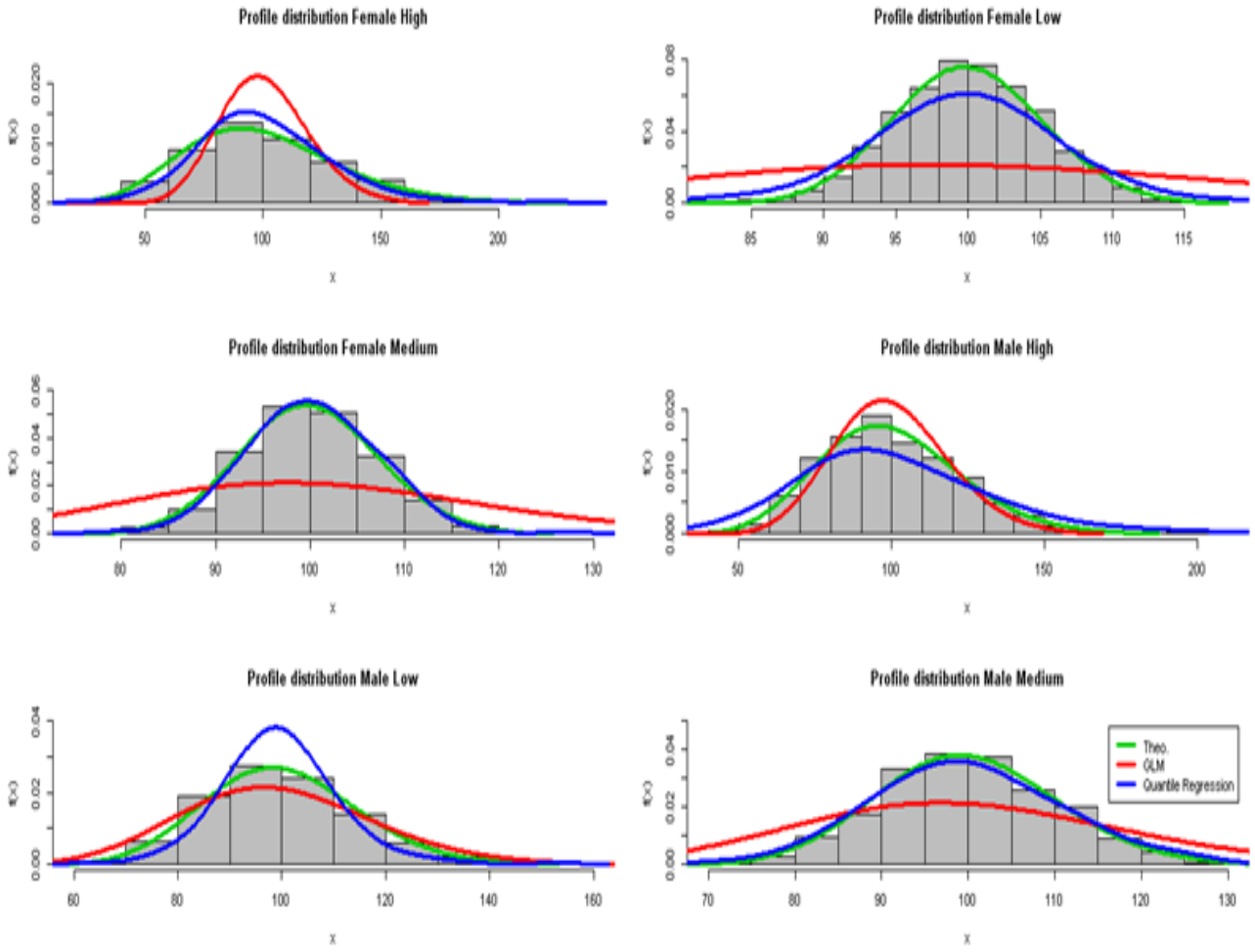
Nel secondo esempio, rappresentato in Figura 12, il campione è stato generato da distribuzioni gamma differenti per ogni profilo $\Gamma(a_{ij}, b_{ij})$.

Si riporta in Tabella 3 la ripartizione dei rischi nelle 6 classi di rischio del nuovo esempio:

Tabella 3: distribuzione di frequenza sesso età

	A	B	M	Totale complessivo
F	570	570	570	**1710**
M	430	430	430	**1290**
Totale complessivo	**1000**	**1000**	**1000**	**3000**

Figura 12: Distribuzioni di probabilità per profilo, in cui il campione di partenza è generato da gamma diverse per ogni profilo



Si osserva, rispetto al caso precedente, che le distribuzioni di probabilità per profilo presentano ovviamente caratteristiche diverse sia in termini di media che in termini di dispersione ed è chiaro che tali diversità, ad un'analisi visiva, sono colte molto meglio dalla Quantile Regression rispetto al GLM gamma. Ad esempio per quanto riguarda il profilo sesso-femmina, età-media, la stima Quantile Regression è evidentemente migliore.

La mancata bontà di adattamento del GLM gamma su alcuni profili è dovuta al fatto che le stime dei parametri sono effettuate nell'ipotesi di parametro di forma costante. Si deduce, dagli esempi sopra riportati, che il GLM ha un buon funzionamento nella stima della distribuzione di probabilità per profilo, solo se le distribuzioni per profilo che generano i dati sono tutte gamma che differiscono per il solo parametro di scala. Si sottolinea, dunque, che il solo passaggio da un campione simulato da un'unica variabile gamma per tutti i profili, a uno simulato da variabili gamma con parametro di forma diverso per profilo, introduce una criticità importante per il GLM gamma nello stimare la distribuzione di probabilità di Z_{ij}.

Poiché finora si è operato considerando variabili aleatorie sempre di tipo gamma e modelli GLM di tipo gamma, il passaggio successivo è quello di vedere come si comportano le due procedure, considerando un database reale di sinistri e quindi non effettuando ipotesi sulla forma della distribuzione che genera i dati.

Il database considerato contiene 44.866 sinistri sui quali sono stati studiati due fattori di rischio:

- età: utilizzata come variabile dicotomica con modalità Giovane-Esperto, a seconda che l'assicurato abbia più o meno di 45 anni;
- tipo: che sintetizza alcune caratteristiche del veicolo, le cui modalità verranno definite con tipo A, tipo B e tipo C.

Nelle seguenti tabelle si riportano alcune statistiche descrittive.

Tabella 4: Distribuzione di frequenza per profilo

Sinistri	Giovane	Esperto	Totale
Tipo A	18,04%	18,31%	7,86
Tipo B	9,08%	10,73%	19,81%
Tipo C	22,64%	21,20%	43,84%
Totale	**49,76%**	**50,24%**	**100,00%**

Tabella 5 Costi medi per profilo

Media	Giovane	Esperto
Tipo A	2637	2273
Tipo B	2381	2093
Tipo C	2280	2063

Tabella 6 Deviazione standard dell'importo del danno

Dev.St	Giovane	Esperto
Tipo A	3674	4857
Tipo B	3878	4423
Tipo C	7691	5779

Tabella 7 75-esimo quantile dell'importo del danno

Quantile 75	Giovane	Esperto
Tipo A	3032	2500
Tipo B	2611	2330
Tipo C	2353	2159

Su questa base dati si sono eseguiti:

- un GLM con ipotesi Gamma
- una Quantile Regression a livello di probabilità $\theta = 0{,}75$.

In Tabella 8 è possibile osservare i parametri stimati ed il loro livello di significatività, constatando che sono tutti significativi.

Tabella 8: Output GLM e QR

Parametri	Coeff QR	P-value QR	Coeff GLM	P-value GLM
Intercetta	8	<10^(-10)	7,86	<10^(-10)
Tipo B	-0,12	<10^(-10)	-0,09	<10^(-10)
Tipo C	-0,21	<10^(-10)	-0,12	<10^(-10)
Esperto	-0,14	<10^(-10)	-0,12	<10^(-10)

In Tabella 9, invece, si riporta, per tutti i profili, un confronto tra le statistiche osservate e gli output dei modelli.

Tabella 9: Confronto tra modelli GLM, Quantile Regression e valori osservati

	Media Osservata		Media Fittata GLM		75-mo quantile osservato		75-mo quantile fittato QR		75-mo quantile fittato GLM	
Modalità	**Giovane**	**Esperto**	**Giovane**	**Esperto**	**Giovane**	**Esperto**	**Giovane**	**Esperto**	**Giovane**	**Esperto**
Tipo A	2637	2273	2604	2301	3032	2500	2974	2586	3609	3190
Tipo B	2381	2093	2374	2098	2611	2330	2645	2300	3290	2908
Tipo C	2280	2063	2306	2038	2353	2159	2415	2100	3196	2824

Si vede immediatamente che:

- il GLM gamma presenta un buon fitting in media: ad esempio per il profilo TipoB-Esperto, la media osservata è 2.093, mentre la stima GLM è 2.098 (errore del 0,3%);
- il 75-esimo quantile osservato, per lo stesso profilo, ammonta a 2.330 e la Quantile Regression presenta un ottimo fitting assestandosi su un valore di 2.300 (errore del -1,2%);
- il fitting del GLM sul 75-esimo percentile non è dei migliori, in quanto il valore fittato ammonta a 2.908, che è sensibilmente diverso rispetto al valore osservato (errore del 25%).

Si osserva, quindi, lo stesso fenomeno che risultava evidente dalla Figura 12, ovvero le difficoltà del GLM gamma nella stima dei quantili per profilo, a differenza dell'ottima performance del modello di Quantile Regression.

Nella

Figura 13, Figura 14 e Figura 15 sono riportati i valori dei coefficienti stimati con il modello QR al variare del livello di probabilità θ, posti a confronto con i coefficienti GLM.

Figura 13 Grafico coefficiente intercetta

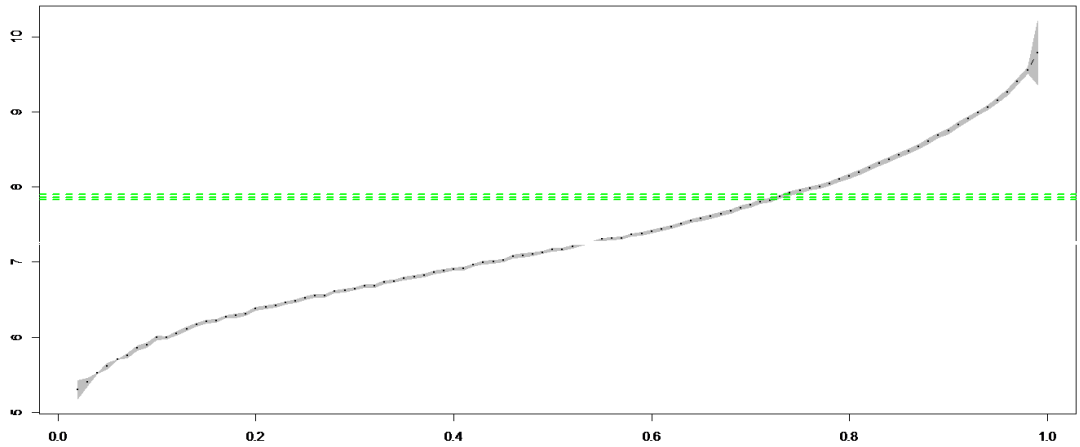


Figura 14: Grafico coefficiente Tipo B

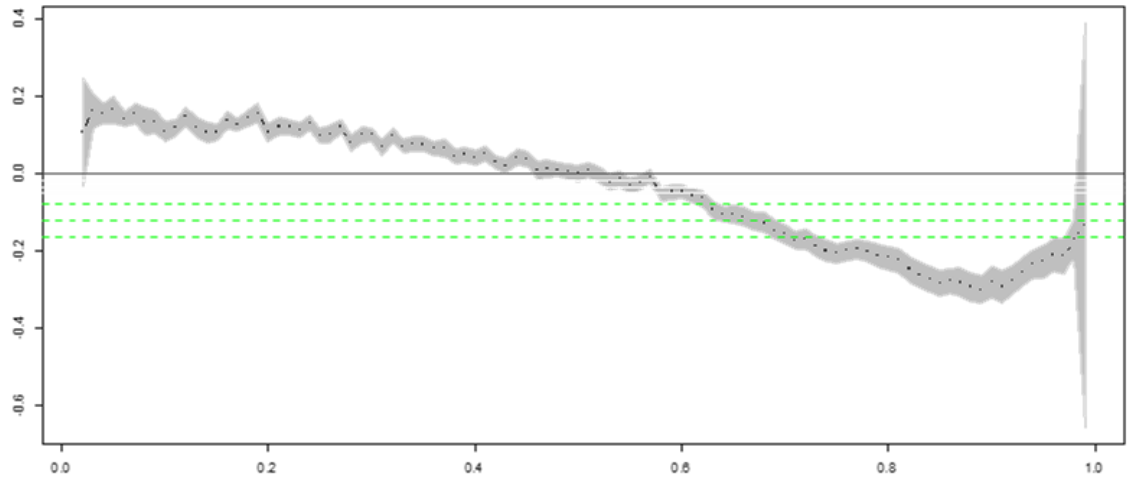
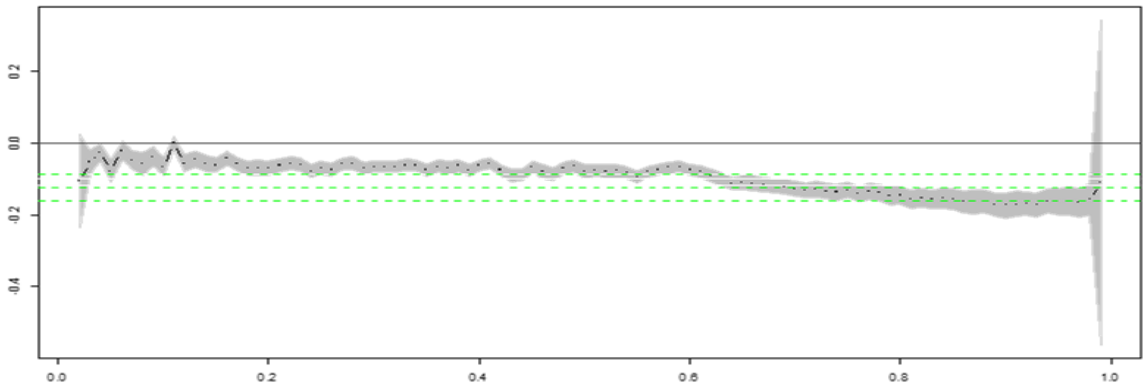


Figura 15: Grafico coefficiente esperto



- I punti neri rappresentano le stime dei parametri ottenute con il modello Quantile Regression (per esempio in
- Figura 13 possiamo notare che in corrispondenza di un livello di probabilità del 40%, otteniamo una stima dell'intercetta approssimativamente uguale a 7),
- la banda grigia rappresenta la stima intervallare al 95%, ottenuta attraverso il metodo del rank test (paragrafo 4.5.2),
- la linea verde centrale tratteggiata rappresenta invece le stime ottenute attraverso il GLM Gamma,
- le linee verdi laterali tratteggiate rappresentano intervalli di confidenza per i parametri del GLM Gamma al 95%, che sono chiaramente costanti al variare del livello di probabilità.

Si nota, che il coefficiente Esperto non presenta cambiamenti rilevanti al variare del livello di probabilità, quindi, in questo caso, una stima con GLM o con QR è abbastanza indifferente; mentre, al contrario, il coefficiente Tipo B è fortemente dipendente dal livello di probabilità. Tuttavia, fissando il livello di confidenza, le stime ottenute con la QR risultano quasi sempre meno variabili, dato che la banda grigia è significativamente più piccola rispetto a quella definita dalle linee verdi del GLM. Quest'ultimo fenomeno sarà tanto più evidente quanto più la distribuzione di probabilità è "densa" nei dintorni del quantile, ovvero dove la sparsity function è piccola (paragrafo 4.5.1). Si vede infatti che avvicinandoci a livelli di probabilità estremi la banda grigia aumenta molto, proprio per la rarefazione dei dati intorno al quantile da stimare.

In conclusione questa prima parte dell'applicazione, limitando l'attenzione a variabili assolutamente continue (come l'importo del danno associato ad un sinistro), permette di sottolineate le difficoltà del GLM gamma nella stima della distribuzione di probabilità per profilo.

6.2. MODELLO A DUE PARTI VS. MODELLO GLM

In questo paragrafo si mostrerà un confronto su un DB realistico, desunto dall'esperienza di una compagnia danni italiana sul ramo RCA, del calcolo del premio effettuato con GLM e con il modello a due parti in base a quanto descritto nel paragrafo (5.3).

Su questo DB sono stati analizzati due fattori di rischio:

- una variabile "tipo guida" che sintetizza in 3 modalità (N, S, X) l'abilità del conducente,
- una variabile "flag convenzione", che sintetizza in 3 modalità (0,1,2), il tipo di convenzione a cui l'assicurato ha avuto accesso.

Si riportano in Tabella 10 e Tabella 11 alcune statistiche descrittive osservate sul database e i relativi indicatori di sintesi.

Tabella 10: valori osservati sul DB di riferimento

Profilo	Numero Rischi	Numero Sinistri	Costo Sinistri
N0	157.119	5.833	24.940.478
N1	136.607	3.370	14.493.182
N2	236.581	5.055	17.607.635
S0	38.011	1.821	8.826.208
S1	29.779	966	8.350.502
S2	65.237	1.790	8.992.558
X0	212.108	9.971	43.696.101
X1	47.614	1.260	3.663.792
X2	266.707	7.379	23.463.252
Totale	**1.189.763**	**37.445**	**154.033.708**

Tabella 11: Indici di sintesi osservati sul DB di riferimento

Profilo	Frequenza Sinistri	Costo Medio	Quota Danni
N0	3,71%	4.276	159
N1	2,47%	4.301	106
N2	2,14%	3.483	74
S0	4,79%	4.847	232
S1	3,24%	8.644	280
S2	2,74%	5.024	138
X0	4,70%	4.382	206
X1	2,65%	2.908	77
X2	2,77%	3.180	88
Totale	**3,15%**	**4.114**	**129**

In Figura 16 e Figura 17 sono riportati i grafici relativi alla Tabella 11

Figura 16: frequenza sinistri osservata per profilo

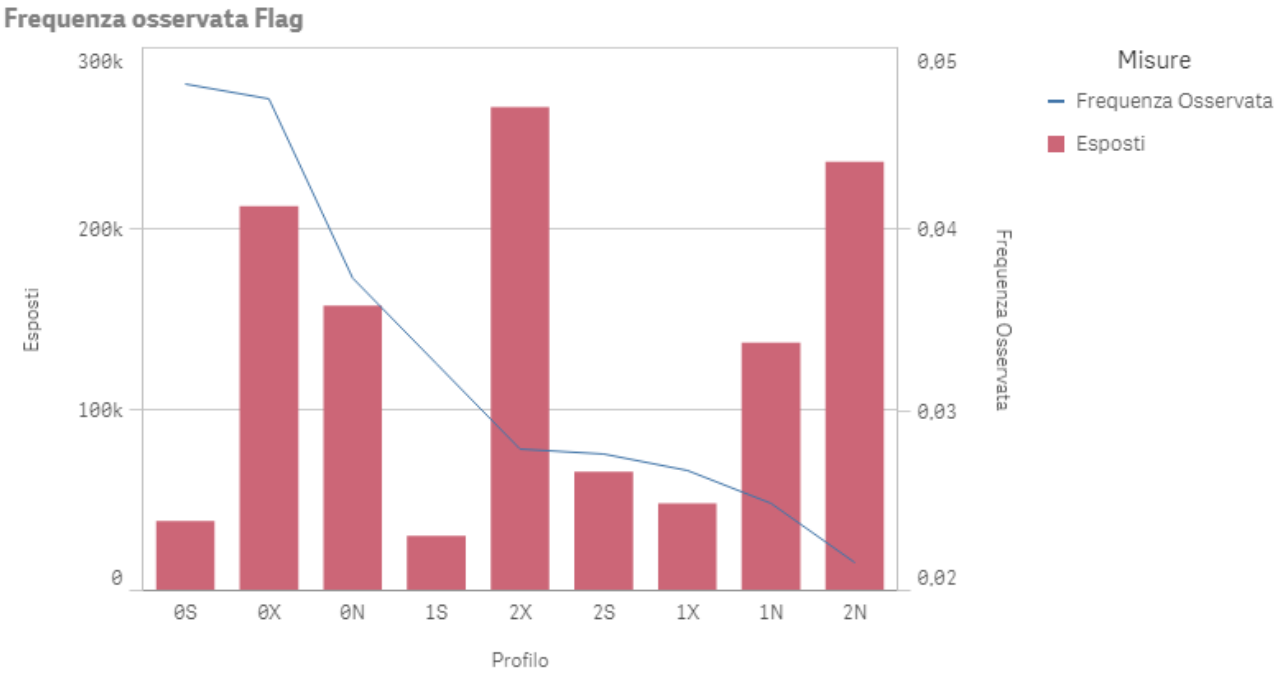
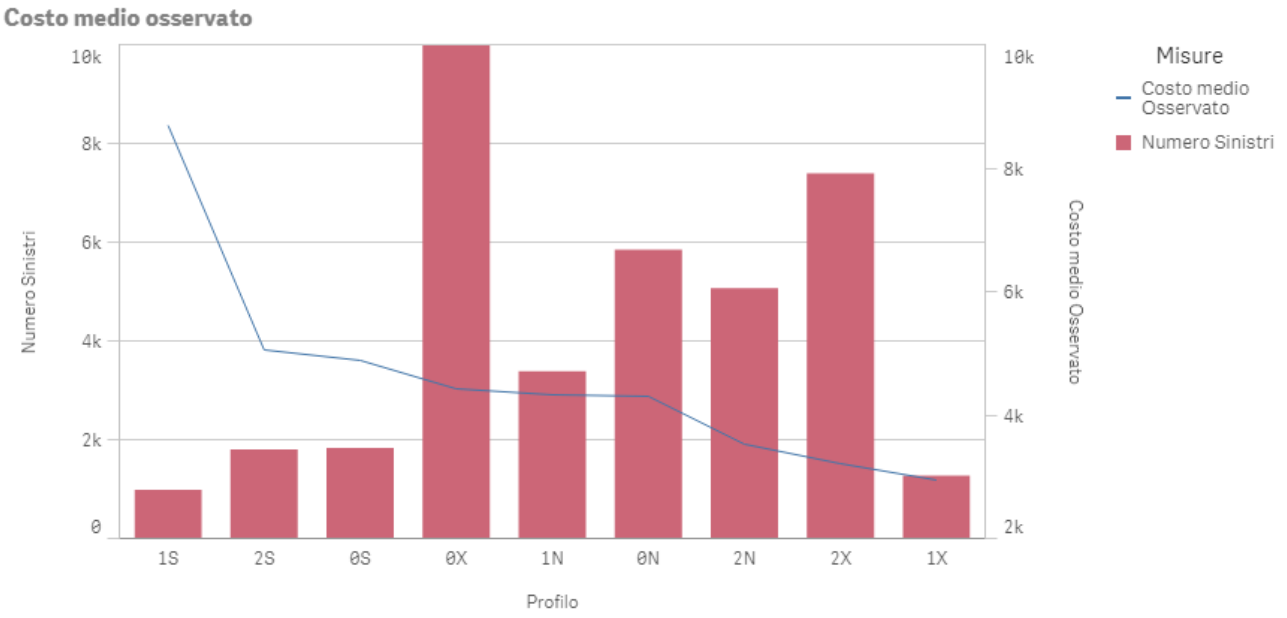


Figura 17: costo medio osservato per profilo



Si vuole introdurre un confronto tra:

- il premio stimato con il GLM (di seguito **premio GLM standard**), sfruttando la relazione $E[Y] = E[N] \cdot E[Z]$:
- il premio stimato con il modello a due parti precedentemente descritto;
- il premio stimato con il GLM, facendo un'analisi separata tra sinistri attritional e large (di seguito **premio GLM attritional-large**)

Prima di procedere al calcolo dei premi occorre fissare il membro di sinistra della FIE, ovvero il fabbisogno puro. Si fissa il fabbisogno puro $Quant_\theta[\ddot{Y}]$ uguale al valore dell'esborso complessivo (Tabella 11) osservato aumentato del 5%. Risulta dunque:

$$Quant_\theta[\ddot{Y}] = 154.033.708 \cdot (1{,}05) = 161.735.393$$

6.2.1. IL CALCOLO DEL PREMIO CON PROCEDURA GLM STANDARD.

Per quanto riguarda la tariffazione eseguita con GLM, si è costruito un modello di Poisson per la frequenza sinistri e si sono ottenuti i seguenti output

Tabella 12: Output modello Poisson

Parametro	Stima	Std.Error	z value	P_Value
Intercetta	- 3,29	0,01	- 314,73	<10^(-10)
Flag 1	- 0,45	0,02	- 28,59	<10^(-10)
Flag 2	- 0,54	0,01	- 48,03	<10^(-10)
Tipo Guida S	0,26	0,02	15,03	<10^(-10)
Tipo Guida X	0,23	0,01	19,90	<10^(-10)

In Tabella 12 sono riportati gli output del modello Poisson, come si può notare tutti i parametri risultano significativi, poiché i p-value sono prossimi allo zero. Si nota, inoltre, che il parametro flag 0 ha una frequenza sinistri più elevata, rispetto alle modalità 1 e 2; mentre il parametro tipo guida N ha una frequenza sinistri più bassa rispetto alle altre due modalità.

Successivamente si è proceduto con la costruzione di un modello gamma per il costo medio e si sono ottenuti i seguenti output.

Tabella 13: Output modello gamma

Parametro	Stima	Std.Error	z value	P_Value
Intercetta	8,38	0,08	105,41	<10^(-10)
Flag 1	- 0,001	0,12	- 0,09	0,924
Flag 2	- 0,25	0,09	- 2,89	0,004
Tipo Guida S	0,37	0,13	2,85	0,004
Tipo Guida X	- 0,05	0,09	- 0,53	0,593

In Tabella 13 sono riportati gli output del modello gamma, come si può notare le modalità Flag 0 e Flag 1 dovrebbero essere raggruppate, così come le modalità Tipo Guida X ed N. Si nota, inoltre, che il gruppo Flag 0-Flag 1 ha un costo medio più elevato, rispetto alla modalità Flag 2; mentre il gruppo tipo guida N-tipo guida X ha un costo medio più basso rispetto alla modalità Tipo Guida S.

Una volta ottenuta una stima del numero atteso dei sinistri con il GLM Poisson e del costo medio con il GLM Gamma, moltiplicando i risultati ottenuti si ottiene una stima della quota danni attesa per ciascun profilo di rischio, ovvero di $E(Y|\boldsymbol{x}_{ij})$.

Si riporta in Tabella 14 un confronto tra le quote danni osservate e i premi GLM per profilo.

Tabella 14: Confronto tra quota danni osservata e stimata con procedura GLM

Profilo	Quota Danni Osservata	Quota danni GLM	Numero Rischi	Esborso totale ossevato	Esborso totale Stimato
N0	159	163	157.119	24.940.478	25.566.404
N1	106	103	136.607	14.493.182	14.033.637
N2	74	74	236.581	17.607.635	17.480.970
S0	232	303	38.011	8.826.208	11.523.035
S1	280	191	29.779	8.350.502	5.699.701
S2	138	138	65.237	8.992.558	8.980.525
X0	206	195	212.108	43.696.101	41.314.396
X1	77	123	47.614	3.663.792	5.855.570
X2	88	88	266.707	23.463.252	23.590.234
Totale	**129**	**129**	**1.189.763**	**154.033.708**	**154.044.472**

Dall'osservazione della Tabella 14, si colgono delle differenze a livello del singolo profilo, ma a livello aggregato le stime GLM coprono la componente equa del fabbisogno (al netto del caricamento) a meno di un piccolo errore di stima.

Ricordando, però che l'obiettivo dell'impresa è il fabbisogno puro, pari a 161.735.393, risulta chiaro che la stima GLM non permette di coprire tale ammontare. Ricordando le considerazioni introdotte nel paragrafo 5.1.1, questo accade poiché:

$$Quant_{\theta}[\ddot{Y}] \neq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} r_{ij}E(Y|\boldsymbol{x}_{ij}).$$

Per completare il calcolo dei premi con metodologia GLM, si calcola il coefficiente correttivo $\tilde{P}^{(0)}$ definito dalla formula (5.1.4). Si ottiene:

$$\tilde{P}^{(0)} = \tilde{P}^{(0)} = \frac{Quant_{\theta}[\ddot{Y}]}{E[\ddot{Y}]} = \frac{161.735.393}{154.044.472} = 1{,}0499.$$

Come ci si aspettava il coefficiente correttivo nel caso GLM è molto vicino a 1,05 e la piccola differenza è dovuta al leggero scostamento tra il danno totale osservato e quello stimato, riportati nelle ultime due colonne della Tabella 14.

Una volta calcolato $\tilde{P}^{(0)}$ si ottengono i premi per profilo calcolati con metodologia GLM attraverso la formula:

$$P_{ij} = \tilde{P}^{(0)} * E(Y|\boldsymbol{x}_{ij}).$$

In Tabella 15 sono riportati i premi GLM finali, comprensivi di caricamento.

Tabella 15: Ammontare finale dei premi calcolati con metodologia GLM

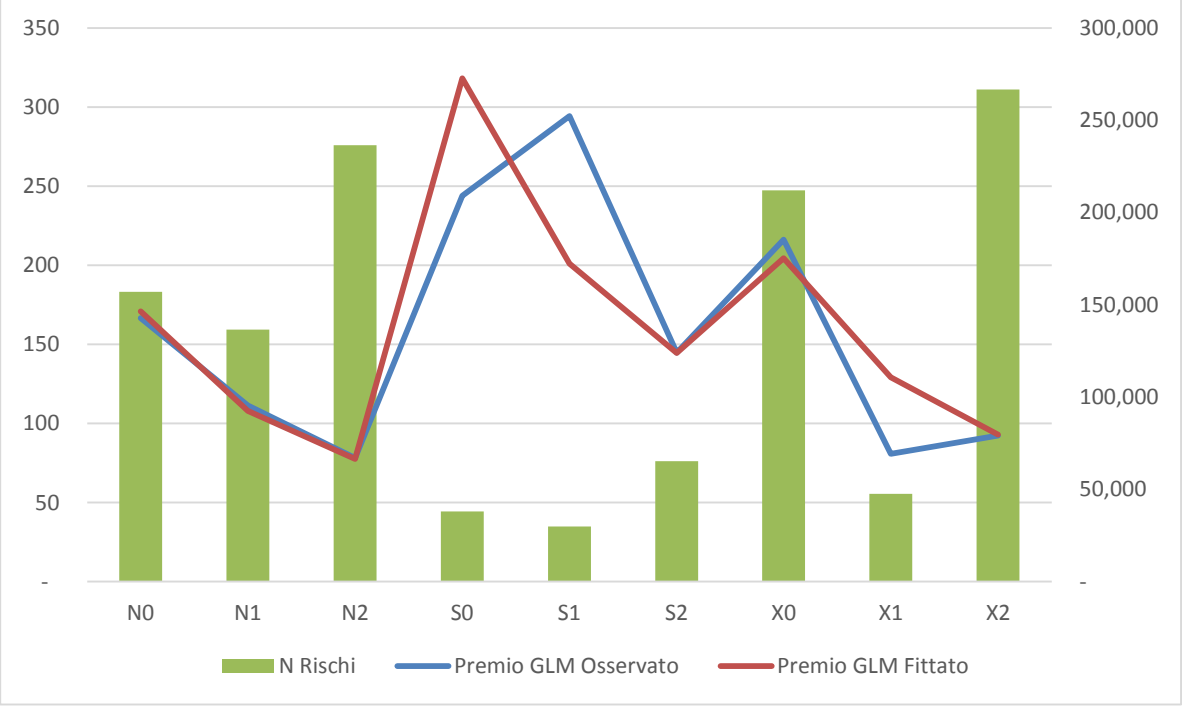
Profilo	$E(Y_{ij})$=Quota danni GLM	$m(Y_{ij})$=Caricamento assoluto	$m(Y_{ij})/E(Y_{ij})$=Caricamento percentuale	$P(Y_{ij})=E(Y_{ij})+m(Y_{ij})$	$N(Y_{ij})$=Numero Rischi	$N(Y_{ij})*P(Y_{ij})$=Esborso totale Stimato
N0	163	8,12	4,99%	171	157.119	26.842.848
N1	103	5,13	4,99%	108	136.607	14.734.289
N2	74	3,69	4,99%	78	236.581	18.353.736
S0	303	15,14	4,99%	318	38.011	12.098.341
S1	191	9,56	4,99%	201	29.779	5.984.267
S2	138	6,87	4,99%	145	65.237	9.428.893
X0	195	9,72	4,99%	205	212.108	43.377.085
X1	123	6,14	4,99%	129	47.614	6.147.919
X2	88	4,42	4,99%	93	266.707	24.768.015
Totale	**129**	**6,47**	**4,99%**	**136**	**1.189.763**	**161.735.393**

Dall'osservazione della seguente tabella si può notare che:

- i premi rispettano il vincolo posto in essere dalla FIE, infatti l'esborso totale osservabile nell'ultima colonna corrisponde al fabbisogno puro obiettivo dell'impresa;
- il caricamento percentuale è lo stesso per tutti i profili come dimostrato nel paragrafo (5.1).

Ponendo l'attenzione sulla bontà di adattamento del modello, si riporta in Figura 18, il grafico che rappresenta il confronto tra premi puri osservati e premi puri stimati con metodologia GLM e che è esplicativo di quanto riportato in Tabella 14 (Rispetto ai dati in Tabella 14, tali premi sono comprensivi della componente di caricamento, e sono ottenuti moltiplicando i dati in tabella per il coefficiente correttivo $\tilde{P}^{(0)}$).

Figura 18: confronto tra quota danni media osservata e stimata con metodologia GLM



Si osserva che su alcuni profili l'adattamento non è dei più soddisfacenti, tale criticità è dovuta a quanto detto nel paragrafo 3.4.2, ovvero che alcuni sinistri, di importo particolarmente elevato, compromettono la

bontà di adattamento del modello. Nel seguito, per superare questa problematica, (paragrafo 6.2.3) verranno mostrati i risultati della procedura GLM attritional-large, che calcola il premio attraverso uno studio separato dei sinistri attritional e dei sinistri large.

6.2.2. IL CALCOLO DEL PREMIO TRAMITE MODELLO A DUE PARTI

Al fine di definire il modello a due parti, si è proceduto in un primo momento alla definizione del modello logit per la stima della probabilità di effettuare almeno un sinistro.

Tabella 16: Output modello logit

Parametro	Stima	Std.Error	z value	P_Value
Intercetta	- 3,34	0,01	- 302,91	<10^(-10)
Flag 1	- 0,42	0,02	- 25,37	<10^(-10)
Flag 2	- 0,50	0,01	- 42,49	<10^(-10)
Tipo Guida S	0,26	0,02	14,79	<10^(-10)
Tipo Guida X	0,23	0,01	19,34	<10^(-10)

In Tabella 16 sono riportati gli output del modello logit, come si può notare tutti i parametri risultano significativi, poiché i p-value sono prossimi allo zero. Si nota, inoltre, che il parametro flag 0 ha una probabilità più elevata di effettuare almeno un sinistro, rispetto alle modalità 1 e 2; mentre il parametro tipo guida N ha una probabilità più bassa di effettuare almeno un sinistro rispetto alle altre due modalità.

Il passo successivo è quello di effettuare una stima del livello di probabilità ottimo θ^*, attraverso la FIE. A tale scopo occorre definire un livello di tolleranza $\varepsilon = 0{,}005$, attraverso il quale applicare l'algoritmo definito nel paragrafo (5.4).

Applicando l'algoritmo si ottiene:

$\theta^* = 0{,}8403$.

Il fatto che il valore di θ^* non sia eccessivamente elevato, definirà delle stime abbastanza efficienti in quanto i quantili da stimare dovrebbero assestarsi in punti della distribuzione non troppo rarefatti (sparsity function piccola). Si è eseguita un analisi di sensitività di θ^* al variare del livello di caricamento scelto per il fabbisogno puro: nel caso in cui si introducesse un caricamento percentuale del 10% invece che del 5%, il valore di θ^* sarebbe pari a 0,8493, quindi molto vicino a quello stimato con la metà del caricamento. Con un caricamento percentuale del 15% si otterrebbe un θ^* pari a 0,8567, ancora una volta molto vicino ai valori stimati con i caricamenti precedenti.

Questa stabilità è dovuta anche all'ampiezza del campione che è pari 1.189.763 rischi. Un possibile sviluppo successivo della tesi potrebbe essere quello di studiare la sensibilità di θ^* al variare del caricamento scelto per il fabbisogno puro, in funzione della numerosità campionaria.

In Tabella 17 è possibile osservare gli output della regressione quantilica.

Tabella 17: Output Quantile Regression 0,8403

Parametro	Stima	Std.Error	z value	P_Value
Intercetta	8,62	0,03	414,15	<10^(-10)
Flag 1	- 0,22	0,03	- 7,15	<10^(-10)
Flag 2	- 0,38	0,02	- 17,12	<10^(-10)
Tipo Guida S	0,09	0,04	2,59	0,010
Tipo Guida X	0,07	0,02	- 3,10	0,002

Come si può notare tutti i parametri risultano significativi, poiché i p-value sono prossimi allo zero; il percentile a livello 0,8403 del danno associato ad un sinistro è più elevato per il flag 0 rispetto all'1 e al 2; mentre il percentile a livello 0,8403 è, per la modalità N, minore di quello della modalità S, ma maggiore di quello della X.

In Tabella 18 sono riportati i premi calcolati con il modello a due parti:

Tabella 18: Premio calcolato tramite modello a due parti

Profilo	Premio 2 parti	Numero Rischi	Esborso totale Stimato
N0	190	157.119	29.813.330
N1	102	136.607	13.898.396
N2	79	236.581	18.772.702
S0	268	38.011	10.172.504
S1	144	29.779	4.287.580
S2	112	65.237	7.329.377
X0	221	212.108	46.962.832
X1	119	47.614	5.668.923
X2	93	266.707	24.777.080
Totale	**143**	**1.189.763**	**161.682.725**

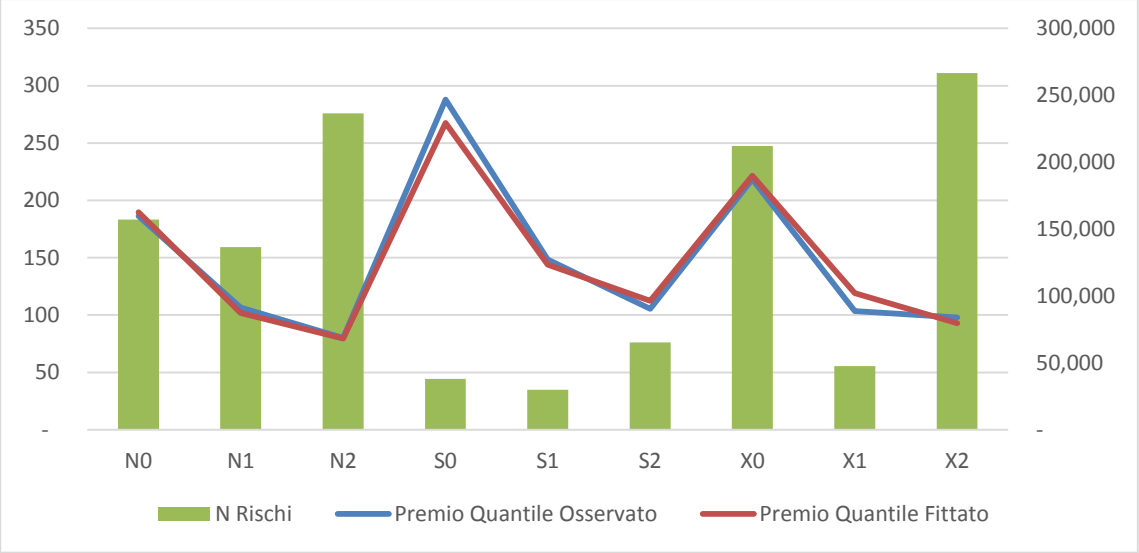
Dall'osservazione di questa tabella si osserva che:

- Il quantile ottimo θ^* fa sì che i premi risultanti da questa procedura, permettano la copertura del fabbisogno puro obiettivo dell'impresa (Figura 19), a meno di una piccola differenza rappresentata dalla tolleranza dell'algoritmo. Infatti l'esborso complessivo stimato è pari a circa 161.682.725, contro il valore di 161.735.393 del fabbisogno puro. (I circa 50.000 euro di differenza sono stati distribuiti in modo proporzionale su tutti i profili)

- In tal caso non è possibile definire in maniera esplicita il caricamento ottenuto, poiché la procedura fornisce in output direttamente il premio puro.

In Figura 19 è riportato il grafico di confronto tra il premio puro osservato e il premio puro calcolato con modello a due parti

Figura 19: confronto tra premio puro osservato e stimato con modello a due parti



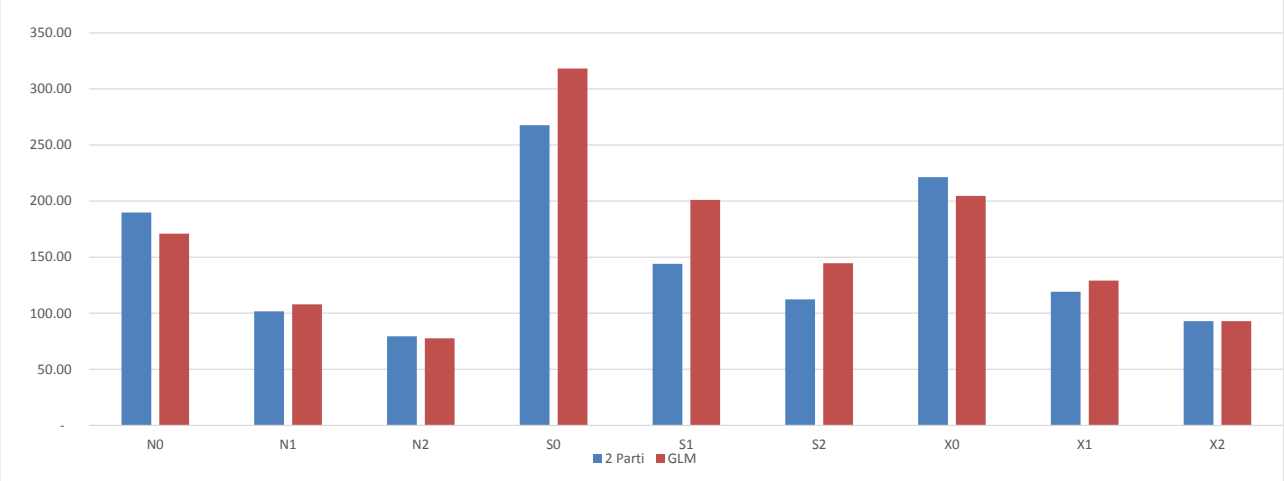
Confrontando i grafici in Figura 19 e Figura 18, si nota immediatamente il miglioramento, in termini di bontà di adattamento, che si ottiene nel passare dalla tecnica GLM a quella a due parti.

In Tabella 19 e Figura 20 sono riportati i premi stimati per profilo con le due procedure e, come si vede, i due criteri producono risultati molto diversi soprattutto su alcuni profili.

Tabella 19: Premi stimati con modello GLM e modello a due parti (aggiungere n rischi e far vedere che torna la FIE)

FLAG_GUIDA	FLAG_CONV	GLM	2 parti	Differenza percentuale
N	0	171	190	11.06%
N	1	108	102	-5.68%
N	2	78	79	2.28%
S	0	318	268	-15.92%
S	1	201	144	-28.36%
S	2	145	112	-22.27%
X	0	205	221	8.26%
X	1	129	119	-7.80%
X	2	93	93	0.03%

Figura 20: Premi stimati con modello GLM e modello a due parti



6.2.3. LA PROCEDURA GLM CON DISTINZIONE TRA SINISTRI ATTRITIONAL E LARGE.

Per ovviare al problema creato dalla presenza di sinistri large che compromettono la bontà di adattamento del GLM standard, si sono suddivisi i sinistri tra attritional e large fissando una soglia a 50.000 euro, (corrispondente al percentile con livello di probabilità 0,99 della distribuzione del costo per sinistro) e si è definito un modello di scomposizione della quota danni (3.4.2) che per comodità riportiamo.

$$E[Y] = E[N]E[Z] = E[N] * E[Z|Z \leq 50.000] * P[Z \leq 50.000] + E[N] * E[Z|Z > 50.000] * P[Z > 50.000]$$

Si è effettuata una stima della quota danni per profilo considerando solo i sinistri attritional (primo addendo della (3.4.2)), ovvero definendo:

- un GLM di Poisson per modellare $E[N]$;
- un GLM gamma per modellare il costo dei sinistri attritional, ovvero $E[Z|Z \leq 50.000]$;
- una stima sui dati osservati e non diversificata per $P[Z \leq 50.000]$.

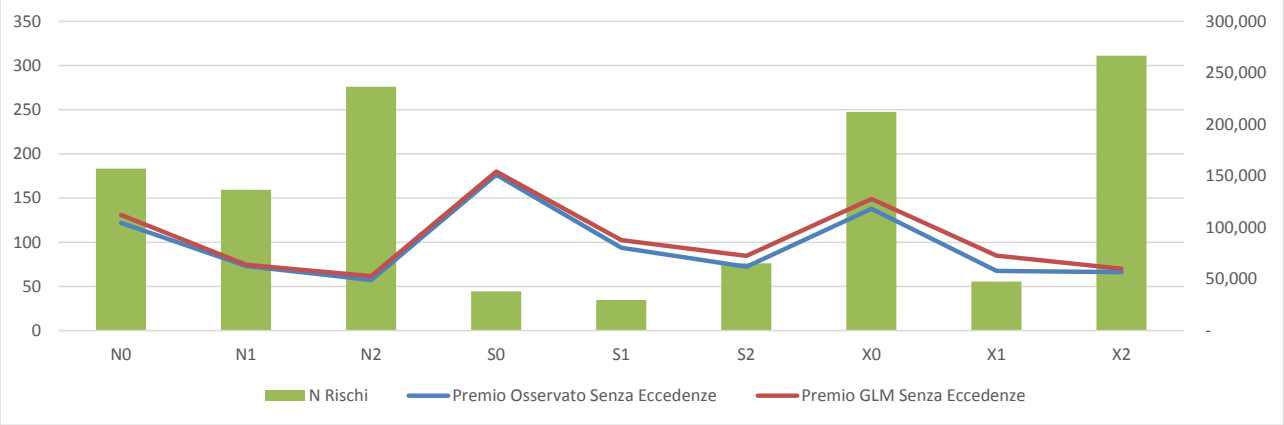
Chiaramente gli output del modello di frequenza sono gli stessi che si vedono in Tabella 12, mentre sono riportati in Tabella 20 gli output del modello di costo medio, condizionato a sinistri di importo inferiore a 50.000 euro.

Tabella 20: Output modello gamma sinistri attritional

Parametro	Stima	Std.Error	z value	P_Value
Intercetta	8,16	0,02	475,13	<10^(-10)
Flag 1	- 0,11	0,03	- 4,45	<10^(-10)
Flag 2	- 0,21	0,02	- 11,44	<10^(-10)
Tipo Guida S	0,06	0,03	2,24	0,030
Tipo Guida X	- 0,10	0,02	- 5,13	<10^(-10)

In (Figura 21) si riporta il confronto tra premi osservati e premi fittati per i sinistri attritional.

Figura 21: Confronto tra quota danni osservata e stimata, sinistri attritional



Dal confronto tra questo grafico e quello in Figura 18, si può evincere quanto i sinistri punta possano pregiudicare il fitting del modello.

Una volta ottenuta la quota danni attritional $E[N] * E[Z|Z \leq 50.000] * P[Z \leq 50.000]$, al fine di definire la quota danni totale, la componente large $E[N] * E[Z|Z > 50.000] * P[Z > 50.000]$ è stata distribuita in maniera moltiplicativa su tutti gli assicurati, ovvero si è individuato un coefficiente α tale che:

$$\alpha = \frac{Costo\ totale\ osservato - Costo\ totale\ attritional\ osservato}{Costo\ totale\ attritional\ osservato} = 34{,}01\%.$$

Si è individuata la quota danni media come:

$$E[Y|\boldsymbol{x}_k] = E[N] * E[Z|\boldsymbol{x}_k \cap Z \leq 50.000] * P[Z \leq 50.000] * (1 + \alpha)$$

Si riportano in Tabella 21 i risultati di tali stime:

Tabella 21: Confronto tra quota danni osservata e stimata con procedura GLM attritional-large

Profilo	Quota Danni Osservata	Quota danni GLM	Numero Rischi	Esborso totale ossevato	Esborso totale Stimato
N0	159	175,43	157.119	24.940.478	27.563.878
N1	106	99,85	136.607	14.493.182	13.640.375
N2	74	82,58	236.581	17.607.635	19.536.211
S0	232	241,07	38.011	8.826.208	9.163.367
S1	280	137,21	29.779	8.350.502	4.086.076
S2	138	113,47	65.237	8.992.558	7.402.226
X0	206	199,82	212.108	43.696.101	42.384.065
X1	77	113,73	47.614	3.663.792	5.415.364
X2	88	94,05	266.707	23.463.252	25.083.409
Totale	**129**	**130**	**1.189.763**	**154.033.708**	**154.274.970**

Dall'osservazione di questi dati si potrebbe desumere un cattivo adattamento dei premi stimati ai valori osservati; tuttavia tale confronto non è sensato in quanto, in questo caso, il modello di regressione è effettuato solo sui sinistri attritional, redistribuendo i sinistri large in maniera identica su tutti i profili.

L'unico confronto attendibile in questo caso è quello rappresentato in Figura 21, dove si confrontano la quota danni attritional osservata e stimata, con risultati soddisfacenti.

Una volta ottenuta una stima di $E[Y|\boldsymbol{x}_k]$, ripercorrendo i passaggi già definiti nel paragrafo 6.2.1 si è ottenuto il coefficiente correttivo $\tilde{P}^{(0)} = 1{,}0484$. Risulta chiaro che gli scostamenti rispetto a 1,05 sono da imputarsi alle differenze tra il danno totale osservato e quello stimato (ultime due colonne della Tabella 21)

In Tabella 22 si riportano i premi stimati con procedura GLM attritional-large.

Tabella 22: Premio calcolato tramite GLM attritional-large

Profilo	$E(Y_{ij})$=Quota danni GLM	$m(Y_{ij})$=Caricamento assoluto	$m(Y_{ij})/E(Y_{ij})$=Caricamento percentuale	$P(Y_{ij})=E(Y_{ij})+m(Y_{ij})$	$N(Y_{ij})$=Numero Rischi	$N(Y_{ij})*P(Y_{ij})$=Esborso totale Stimato
N0	175,43	8,48	4,84%	184	157.119	28.896.811
N1	99,85	4,83	4,84%	105	136.607	14.299.995
N2	82,58	3,99	4,84%	87	236.581	20.480.942
S0	241,07	11,66	4,84%	253	38.011	9.606.489
S1	137,21	6,64	4,84%	144	29.779	4.283.670
S2	113,47	5,49	4,84%	119	65.237	7.760.182
X0	199,82	9,66	4,84%	209	212.108	44.433.671
X1	113,73	5,50	4,84%	119	47.614	5.677.240
X2	94,05	4,55	4,84%	99	266.707	26.296.392
Totale	**130**	**6,48**	**4,84%**	**136**	**1.189.763**	**161.735.393**

Dall'osservazione della tabella si può notare che:

- i premi rispettano il vincolo posto in essere dalla FIE, infatti l'esborso totale osservabile nell'ultima colonna corrisponde al fabbisogno puro obiettivo dell'impresa;
- il caricamento percentuale è lo stesso per tutti i profili come dimostrato nel paragrafo (5.1).

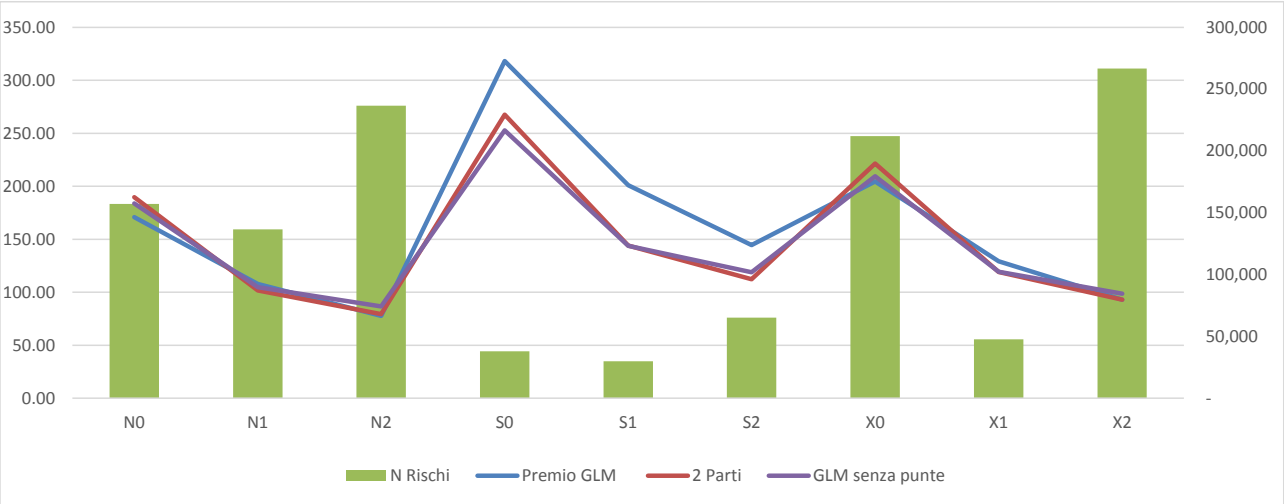
6.2.4. IL CONFRONTO TRA I VARI APPROCCI..

In Tabella 23 e Figura 22 è riportato il confronto tra i premi per profilo calcolati con i 3 metodi:

Tabella 23: Confronto tra i premi calcolati con le tre procedure

PROFILO	GLM	2 Parti	GLM senza punte
N0	170,86	189,75	183,92
N1	107,87	101,74	104,68
N2	77,58	79,35	86,57
S0	318,31	267,62	252,73
S1	200,97	143,98	143,85
S2	144,54	112,35	118,95
X0	204,52	221,41	209,49
X1	129,13	119,06	119,23
X2	92,87	92,90	98,60

Figura 22 confronto tra i premi calcolati con i 3 metodi



Dal confronto tra i tre premi calcolati si può notare che:

- i premi dedotti con modello a due parti e con GLM attritional-large sono abbastanza simili e si discostano dal premio calcolato con GLM standard;
- la procedura GLM frequenza-costo medio, presenta degli output critici da un punto di vista di "goodness of fitting" (Figura 18) e sembra quindi, tra le tre, l'opzione peggiore.

A scopo puramente comparativo, si riportano in Tabella 24 i caricamenti dei premi che si otterrebbero confrontando i premi puri, calcolati con le tre procedure, con il premio equo $E[Y|\boldsymbol{x}_k]$ stimato con la procedura GLM standard (paragrafo 6.2.1).

In Tabella 24, si riportano i livelli di caricamento stimati nei tre approcci, sia in termini assoluti che in percentuale del valore atteso.

Tabella 24: Caricamenti dei premi stimati con i tre metodi sia in termini assoluti che in termini di valore atteso

	Caricamento in termini assoluti			Rapporto caricamento valore atteso		
PROFILO	GLM	2 Parti	GLM attritional-large	GLM	2 Parti	GLM attritional-large
N0	8,14	27,03	21,20	4,99%	16,61%	13,03%
N1	5,14	-0,99	1,95	4,99%	-0,96%	1,90%
N2	3,69	5,46	12,68	4,99%	7,39%	17,16%
S0	15,16	-35,53	-50,42	4,99%	-11,72%	-16,63%
S1	9,57	-47,42	-47,55	4,99%	-24,78%	-24,84%
S2	6,88	-25,31	-18,71	4,99%	-18,39%	-13,59%
X0	9,74	26,63	14,71	4,99%	13,67%	7,55%
X1	6,15	-3,92	-3,75	4,99%	-3,19%	-3,05%
X2	4,42	4,45	10,15	4,99%	5,03%	11,47%

Dall'osservazione di questi risultati si nota che:

- come spiegato nel paragrafo (5.1.1), il caricamento ottenuto con procedura GLM è calcolato in proporzione costante ai rispettivi valori attesi;
- il modello a due parti e il modello GLM-attritional presentano risultati simili come era già evidente dall'osservazione della Tabella 23 e della Figura 22;
- Il GLM attritional-large definisce caricamenti percentuali diversi, perché sono rapportati a una stima equa ottenuta con il modello GLM standard (si otterrebbe un unico valore pari a 4,84% se i caricamenti fossero rapportati alla stima equa ottenuta con modello a due parti);
- il modello a due parti e il modello GLM attritional large, determinano per alcuni profili, un caricamento individuale negativo. Questo significa che, in tali casi, il quantile ottimo determinato è inferiore al valore atteso stimato con procedura GLM standard.

L'ultima analisi condotta riguarda il ranking dei premi: ci si è chiesti se i premi calcolati con le tre metodologie si differenziassero anche in termini di ordinamento.

Confrontando i ranking dei premi per profilo troviamo che essi coincidono nel modello a due parti e nel GLM attritional-large, mentre rispetto al GLM sono invertiti le posizioni 3-4 e 5-6 (Tabella 25).

Tabella 25: ranking dei premi puri per profilo

PROFILO	GLM	2 Parti	GLM senza punte
N0	4	3	3
N1	7	7	7
N2	9	9	9
S0	1	1	1
S1	3	4	4
S2	5	6	6
X0	2	2	2
X1	6	5	5
X2	8	8	8

Poiché la procedura GLM e la procedura GLM attritional-large differiscono per il fatto che nel secondo non si ha una profilazione del rischio per i sinistri large, si imputano le differenze ai sinistri large stessi. Si è concluso che la modalità S è quella in cui i sinistri large hanno avuto maggiore incidenza (anche perché è la meno popolata). Infatti si è costruito un sotto-database con le sole osservazioni uguali a S e confrontando il grafico della quota danni media osservata totale e attritional, si vede che il rango dei profili S0 ed S1 risulta invertito.

Figura 23: Quota danni media osservata, limitata a osservazioni con modalità S

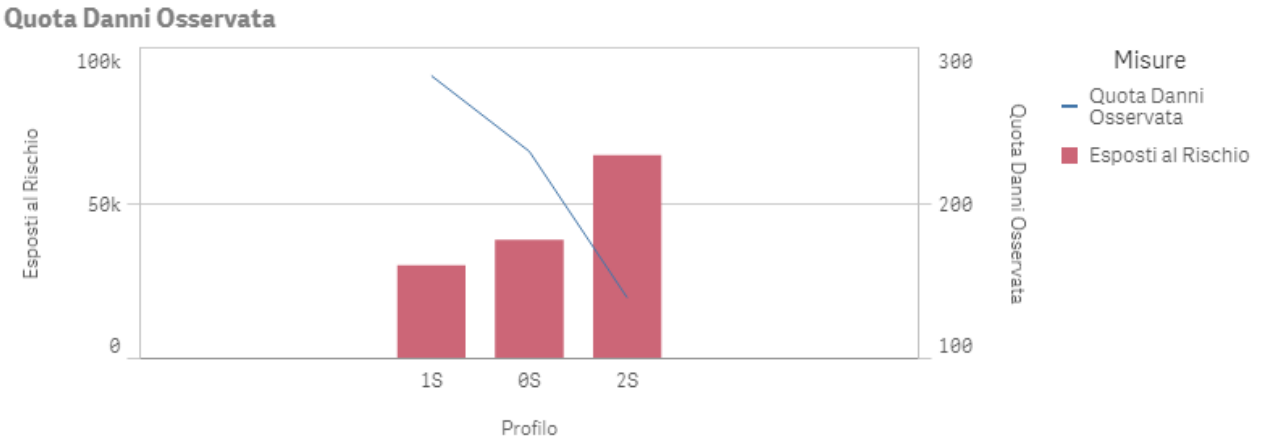
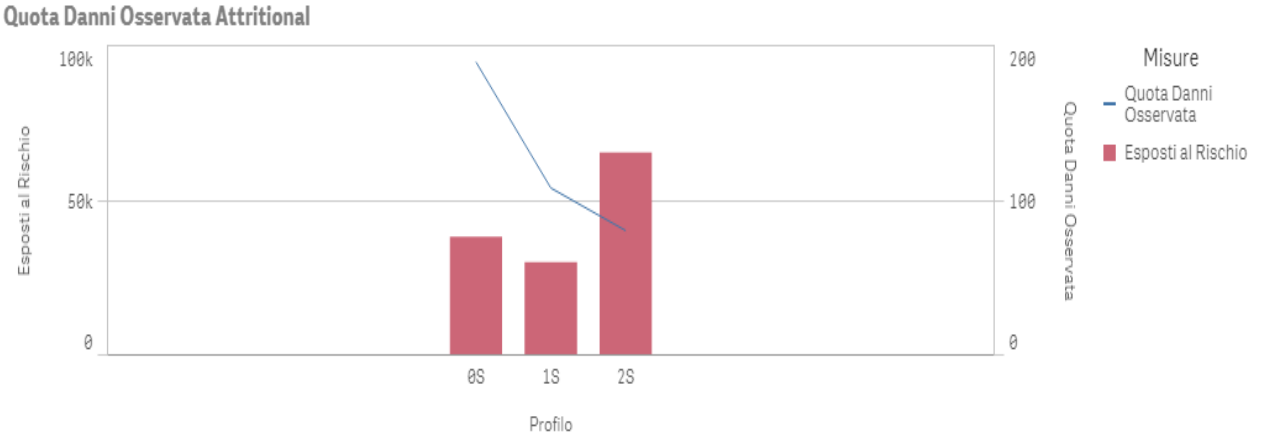


Figura 24: Quota danni media osservata attritional, limitata a osservazioni con modalità S



Si nota immediatamente che, condizionatamente alla modalità S la modalità flag 1 è la più rischiosa; tutavia escludendo i sinistri large dall'analisi, la modalità flag 0 risulta di gran lunga la più rischiosa. Tale cambio di ordinamento è la causa delle differenze osservabili in Tabella 25.

6.3. CONCLUSIONI

Al fine di valutare la qualità della Quantile Regression, nella prima parte dell'applicazione si sono eseguiti dei confronti su variabili risposta continue tra la regressione GLM gamma e quella del quantile. Dai risultati ottenuti su diversi database più o meno favorevoli all'una o all'altra ipotesi, si è concluso che il GLM ha un buon comportamento nella stima del valore atteso della distribuzione per profilo, ma è, nella stragrande maggioranza dei casi, inadeguato per il calcolo dei quantili della medesima distribuzione.

La Quantile Regression non si limita a fornire una buona stima del singolo quantile, ma al variare del livello di probabilità a cui viene eseguita, permette di ottenere in output una stima dell'intera distribuzione di probabilità per profilo, con risultati migliori rispetto alla stima fornita dal GLM.

Si è notato inoltre, costruendo intervalli di confidenza sui parametri stimati, che la QR fornisce delle stime dei parametri meno volatili rispetto alle stime GLM, quasi per ogni livello di probabilità scelto.

Nell'ottica di inserire la Quantile Regression nel contesto della tariffazione, si è costruito un impianto tariffario originale che consenta di determinare un premio individuale in cui il caricamento non sia più determinato in modo proporzionale al valore atteso (come accade nell'impianto GLM utilizzato nella pratica), ma tenendo conto delle caratteristiche individuali degli assicurati. Tale impianto è stato costruito sfruttando un modello a due parti.

Per analizzare i punti di forza e di debolezza della nuova procedura introdotta quest'ultima è stata confrontata con l'impianto tariffario GLM standard, e con un impianto tariffario basato sempre sul GLM, ma che trattasse in maniera diversa i sinistri attritional e i sinitri large.

In primis si è osservato che la procedura GLM standard, presenta una bontà di adattamento non soddisfacente ai dati osservati, proprio perché non trattando separatamente i sinistri large, questi compromettono sul modello di costo medio la bontà di adattamento.

Confrontando la procedura a due parti e quella GLM attritional-large, i risultati in termini di bontà di adattamento risultano notevolmente migliorati e in termini di premi stimati fortemente simili. C'è da sottolineare che il modello a due parti permette di definire il premio senza dover stimare una soglia di separazione tra sinistri attritional e sinistri large, quindi senza introdurre una scelta soggettiva da parte del valutatore.

Infine si è voluto indagare se le tre procedure oltre a fornire risultati diversi in termini assoluti, fornissero risultati diversi anche in termini di ranking di rischiosità dei profili. Si è notato, limitatamente a questo esempio, che i ranking sono uguali considerando la procedura GLM attritional-large e il modello a due parti, mentre sono diversi nel GLM standard. Si è mostrato che la differenza è ancora dovuta ai sinistri large. Nell'ipotesi comunemente accettata che i sinistri large siano indipendenti dalle caratteristiche degli individui, l'utilizzo di una procedura GLM che non tratti separatamente i sinistri attritional e large, oltre a compromettere la bontà di adattamento del modello puo' definire anche un ranking di rischiosità dei profili non coerente. Si sottolinea che la procedura a due parti fornisce, in questo esempio, il medesimo ranking del GLM attritional large, senza fissare una soglia in modo soggettivo, ma semplicemente sfruttando la proprietà di robustezza dei quantili.

Occorre notare, tuttavia, che i premi ottenuti tramite il modello a due parti sono, sia in termini di ranking che in termini assoluti, funzione del quantile ottimo θ^*, che è a sua volta funzione del caricamento $m(\ddot{Y})$ applicato nel membro di sinistra della FIE. Partendo da un diverso valore di $m(\ddot{Y})$, si otterrebbe un diverso valore di θ^*, che potrebbe comportare dei risultati diversi non solo in termini numerici, ma anche in termini di ranking dei premi per profilo (La cosa potrebbe comunque accadere fissando una diversa soglia di separazione tra sinistri attritional e large).

Il rischio a cui si può andare incontro è che su portafogli piccoli, fissando caricamenti $m(\ddot{Y})$ molto onerosi, si potrebbe raggiungere un livello di probabilità ottimo θ^*molto elevato; in tali punti la distribuzione per profilo è molto rarefatta (sparsity function molto grande) e variando anche di poco la scelta di $m(\ddot{Y})$, si potrebbero ottenere risultati profondamente diversi.

In conclusione, al fine di definire un impianto tariffario originale con l'obiettivo di allocare il fabbisogno puro (obiettivo dell'impresa) sulle singole teste assicurate, si è introdotto un modello a due parti in cui la seconda di esse è una quantile regression. Tali premi sono stati confrontati con quelli definiti dalle procedure GLM standard e GLM attritional-large, osservando immediatamente che la prima delle due non presenta un buon adattamento ai dati osservati, proprio a causa dei sinistri large stessi.

La procedura a due parti e quella GLM attritional-large presentano risultati abbastanza simili e abbastanza soddisfacenti dal punto di vista della bontà di adattamento, con la differenza che il modello a due parti non richiede la determinazione di una soglia per separare i sinistri attritional e large.

Con l'introduzione di questo nuovo impianto tariffario si introduce il fatto che il premio pagato dall'individuo è conseguenza diretta della scelta fatta dall'assicuratore nel fissare l'obiettivo dell'impresa, ovvero nel definire il caricamento $m(\ddot{Y})$. I premi sono fortemente influenzati da questa scelta, sia in termini assoluti che in termini di ordinamento ed è chiaro che questa dipendenza possa risultare difficile da gestire. Se da un lato questa caratteristica può essere vista come un problema, da un altro è uno dei punti di forza del nuovo impianto, in quanto a oggi nella pratica attuariale la profilazione del rischio risulta indipendente dalla scelta di $m(\ddot{Y})$.

7. BIBLIOGRAFIA

[1] Artzner P., Delbaen F., Eber J., Heath D. (1999): "Coherent measures of risk", Mathematical Finance, 9, 203-228.

[2] Bassett G. W., Koenker R. (1986): "Strong Consistency of Regression Quantiles and Related Empirical Processes", Econometric Theory, 2, 191-201.

[3] Bickel, P. J. (1976): "Another Look At Robustness; A Review of Reviews and Some New Developments",Scandinavian Journal of Statistics, 3, 145-158.

[4] Bofinger, E. (1975): "Estimation of a Density Function Using Order Statistics", The Australian Journal of Statistics, 17, 1-7.

[5] Davis, R. A., W. T. M. Dunsmuir (1997): "Least Absolute Deviation Estimation for Regression With ARMA Errors", Journal of Theoretical Probability, 10, 481-497.

[6] Davis, R. A., K. Knight, and J. Liu (1992): "M-estimation for Autoregressions With Innite Variance", Stochastic Processes and their Applications, 40, 145-180.

[7] De Jong P., Heller G. (2012): "GLM Insurance data", Cambridge

[7] Duan N., Manning W.G., Morris C.N., Newhouse JP (1983): " A comparison of alternative models for the demand for medical care". J Business Econ Statist. 1983, 1, 115–126.

[9] Frank, I. E., and J. H. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools", Technometrics, 35, 109-135.

[10] Gigante P., Picech L., Sigalotti L. (2010): " La tariffazione nei rami danni con modelli lineari generalizzati", EUT

[11] Gutenbrunner, C., J. Jureckova , R. Koenker, and S. Portnoy (1993): "Tests of linear hypotheses based on regression rank scores", J. of Nonparametric Statistics, 2,307-330.

[12] Gutenbrunner, C., and J. Jureckova (1992): "Regression quantile and regression rank score process in the linear model and derived statistics", Ann. Statist., 20, 305-330.

[13] Hall, P., and S. Sheather (1988): "On the distribution of a studentized quantile", J. Royal Stat. Soc. (B), 50, 381-391.

[14] He, X., and P. Ng (1999): "COBS: qualitatively constrained smoothing via linear programming", Computational Statistics, 14, 315-337.

[15] He, X., and Q.-M. Shao (1996): "A General Bahadur Representation of M-estimators and Its Application to Linear Regression With Nonstochastic Designs", The Annals of Statistics, 24, 2608-2630.

[16] Huber, P. J. (1973): "Robust Regression: Asymptotics, Conjectures and Monte Carlo", The Annals of Statistics, 1, 799-821.

[17] Hudson I., Ford R. P. (2003): “Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS”, Computational Statistics & Data Analysis

[18] Jureckova J. (1999): “Regression rank-scores tests against heavy-tailed alternatives”, Bernoulli, 5.

[19] Jureckova, J., and P. K. Sen (1984): “On Adaptive Scale-equivariant M-estimators in Linear Models”, Statistics & Decisions Supplement Issue, 1, 31-46.

[20] Jureckova, J. (1977): “Asymptotic Relations of M-estimates and R-estimates in Linear Regression Model”, The Annals of Statistics, 5, 464-472.

[21] Jureckova, J., e B. Prochazka (1994): “Regression Quantiles and Trimmed Least Squares Estimator in Nonlinear Regression Model”, Journal of Nonparametric Statistics, 3, 201-222.

[22] Koenker, R., and G. Bassett (1978): “Regression Quantiles”, Econometrica, 46, 33-50.

[23] Koenker, R. , G. W. Bassett (1984): “Four (pathological) Examples in Asymptotic Statistics” The American Statistician, 38, 209-212.

[24] R. Koenker, K. Hallock, (2001): “Quantile Regression”, Journal of Economic Perspective, 15, 143–156

[25] Koenker, R., and Q. Zhao (1994): “L-Estimation for linear heteroscedastic models”, J. of Nonparametric Statistics, 3, 223-235.

[26] Koenker, R., and J. Machado (1999): “Goodness of fitt and related inference processes for Quantile Regression”, J. of Am. Stat. Assoc., 94, 1296-1310.

[27] Koenker, R. (1988): “Asymptotic Theory and Econometric Practice”, Journal of Applied Econo- metrics, 3, 139-147.

[28] Koenker, R. (1994): “Confidence intervals for regression quantiles”, in Asymptotic Statistics: Proceedings of the 5th Prague Symposium on Asymptotic Statistics.

[29] Kudryavtsev A. (2009): “Using Quantile Regression for rate-making”, Insurance: Mathematics and Economics, 45, 296-304

[30] Nelder J.A. and Wedderburn R. W. M. (1972): “Generalized Linear Model”, Journal of the Royal Statistical Society, 135, 370-384

[31] Olivieri A. Pitacco E. (2010):” Introduction to insurance mathematics: technical and financial features of risk transfers”, Springer

[32] Pitacco E. (2000):”Elementi di matematica delle assicurazioni”, Lint Editoriale

[33] Ruppert, D., and R. Carroll (1980): “Trimmed least squares estimation in the linear model”, Journal of the American Statistical Association, 75, 828-838

[34] Sánchez L., Lachos V., Labra F., (2013): "Likelihood Based Inference for Quantile Regression Using the Asymmetric Laplace Distribution", Departamento de Estatística, Universidade Estadual de Campinas, Brazil

[35] Siddiqui, M. (1960): "Distribution of Quantiles from a Bivariate Population", Journal of Research of the National Bureau of Standards, 64, 145-150.

[36] Weiss, A. A. (1991): "Estimating Nonlinear Dynamic Models Using Least Absolute Error Estimation", Econometric Theory, 7, 46-68.

[37] Zhou K. , Portnoy S. (1996): "direct use of regression quantiles to construct confidence sets in linear models", The annals of statistics, 24, 287-306